EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

Instituto Politécnico Nacional
"La Técnica al Servicio de la Patria"

Research in Computing Science

Vol. 153 No. 10
October 2024

Research in Computing Science

***RESEARCH IN COMPUTING SCIENCE**, Año 24, Volumen 153, No. 10, Octubre de 2024, es una publicación mensual, editada por el Instituto Politécnico Nacional, a través del Centro de Investigación en Computación. Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, Ciudad de México, Tel. 57 29 60 00, ext. 56571. https://www.rcs.cic.ipn.mx. Editor responsable: Dr. Grigori Sidorov. Reserva de Derechos al Uso Exclusivo del Título No. 04-2019-082310242100-203. ISSN: en trámite, otorgado por el Instituto Nacional del Derecho de Autor. Responsable de la última actualización de este número: el Centro de Investigación en Computación, Dr. Grigori Sidorov, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738. Fecha de última modificación 04 de Octubre de 2024.*

***RESEARCH IN COMPUTING SCIENCE**, Year 24, Volume 153, No. 10, October, 2024, is a monthly publication edited by the National Polytechnic Institute through the Center for Computing Research. Av. Juan de Dios Bátiz S/N, Esq. Miguel Othón de Mendizábal, Nueva Industrial Vallejo, C.P. 07738, Mexico City, Tel. 57 29 60 00, ext. 56571. https://www.rcs.cic.ipn.mx. Editor in charge: Dr. Grigori Sidorov. Reservation of Exclusive Use Rights of Title No. 04-2019-082310242100-203. ISSN: pending, granted by the National Copyright Institute. Responsible for the latest update of this issue: the Computer Research Center, Dr. Grigori Sidorov, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738. Last modified on October 4, 2024.*

Advances in Pattern Recognition

J. Arturo Olvera-López
J. Francisco Martínez-Trinidad
J. Ariel Carrasco-Ochoa
Efren Mezura-Montes
Hector Gabriel Acosta-Mesa (eds.)

Instituto Politécnico Nacional, Centro de Investigación en Computación
México 2024

ISSN: in process

Formerly ISSNs: 1870-4069, 1665-9899

Instituto Politécnico Nacional (IPN)
Centro de Investigación en Computación (CIC)
Av. Juan de Dios Bátiz s/n esq. M. Othón de Mendizábal
Unidad Profesional "Adolfo López Mateos", Zacatenco
07738, México D.F., México

http://www.rcs.cic.ipn.mx
http://www.ipn.mx
http://www.cic.ipn.mx

Indexed in LATINDEX, DBLP and Periodica

Electronic edition

ISSN 1870-4069

Table of Contents

Page

ISSN 1870-4069

A Copula Functions Approach for Predicting Temporomandibular Disorders

Carlos López-Hernández[1], Rogelio Salinas-Gutiérrez[1], Omar Gutiérrez-Navarro[2], Ivonne Bazán-Trujillo[2]

[1] Universidad Autónoma de Aguascalientes, Centro de Ciencias Básicas, Mexico

[2] Universidad Autónoma de Aguascalientes, Centro de Ciencias de la Ingeniería, Aguascalientes, Mexico

{al223348, rogelio.salinas, omar.gutierrezn, ivonne.bazan }@edu.uaa.mx

Abstract. Temporomandibular disorders (TMD) are a set of conditions affecting the temporomandibular joint and associated muscles. This work addresses the classification of healthy and TMD diagnosed individuals using supervised learning models. 118 features were extracted from mandibular movement signals and a selection of those features is used for training the classification models. The dataset of signals was obtained via a marker tracking system with a depth camera. A set of 5 classifiers including two based on copula functions that model dependence structures is tested. The performances of these classifiers are compared. Results show good performance for the copula classifiers, although they do not show better performance than those of models such as KNN, SVM or Naive Bayes.

Keywords: Temporomandibular disorders, feature dependencies, supervised classification, copula functions.

1 Introduction

Temporomandibular disorders (TMD) are a group of musculoskeletal conditions that involve the temporomandibular joint, the masticatory muscles and all associated tissues. TMD are one of the most common causes of pain in the mouth and face, and have the potential to produce chronic pain [6]. The causes behind TMD are multifactorial, including biological, psychological and social causes. The diagnosis of TMD depends largely on clinical history and physical evaluations such as the DC-TMD of [13].

There are several works addressing the classification of TMD using supervised learning models. They use different kind of data such as sound [15], movement [2], or EMG [16]. With this data, models such as KNN, naive Bayes, or SVM have been trained. However, to the best of our knowledge there has not been attempts

of predicting TMD with copula function models. Copula functions are joint distribution functions that provide a flexible way of modeling association among features. They enable to separate the dependence structure from the marginal distributions of the features[11]. Being able to model both linear and non linear dependencies, copula functions have been used in fields such as economics and finance [5], hydrology [8], evolutionary computation [12], multispectral image processing [17] among others.

This work proposes the use of Frank and Gaussian bivariate copulas for modeling dependencies among features and a tree graphical model for the selection of the strongest association. Probabilistic classifiers based on these copulas are trained in the classification of healthy individuals and those with temporomandibular disorders, aiming for an improvement performance based on the dependency information provided by the copula functions. A Mandibular movement dataset from healthy and TMD diagnosed individuals is used for training the models. This dataset was obtained via a marker tracking system with a depth camera.

The content of the paper is as follows: Section 2 provides a brief introduction to the theory of copula functions with emphasis on the selected copulas and the probabilistic classifier. Section 3 describes data used in this article; Section 4 narrates the methodological process followed. The results of the experiments are found in Section 5, and finally Section 6 concludes the paper summarizing the work done and our findings.

2 Copula Theory

2.1 Copula Functions

Copula functions were first introduced by Sklar [14] as a way to differentiate the effect of dependencies from that of marginal distributions in a joint distribution function. By using copula functions it is possible to model both linear and nonlinear dependencies among features.

Definition 1. *A copula C is a joint distribution function of standard uniform variables $U_1, U_2, ..., U_d$: $C(u_1, ..., u_d) = P(U_1 \leq u_1, ..., U_d \leq u_d)$ where each variable follows a standard uniform distribution $U_i \sim Uniform(0, 1)$ for $i = 1, 2, ..., d$.*

Sklar's Theorem states how the copula function relates to any joint distribution (see [14]). As a consequence of this, as shown in Equation (1) any d-dimensional joint density f and their marginal densities $f_1, f_2, ..., f_d$ are also related and can be represented as:

$$f(x_1, x_2, ..., x_d) = c(F_1(x_1), F_2(x_2), ..., F_d(x_d)) \cdot \prod_{i=1}^{d} f_i(x_i), \tag{1}$$

where c is the density of the copula C, X_i is a continous random variable, $F_i(x_i)$ is the marginal distribution function of X_i and $f_i(x_i)$ is the marginal density of

Table 1. Copula Density Functions.

Copula	Density Function	Parameter
Frank	$c(u_1, u_2, \theta) = \frac{-\theta(e^{-\theta}-1)e^{-\theta(u_1+u_2)}}{((e^{-\theta u_1}-1)(e^{-\theta u_2}-1)+(e^{-\theta}-1))^2}$	$\theta \in (-\infty, \infty) - \{0\}$
Gaussian	$c(u_1, u_2, \theta) = (1-\theta^2)^{-\frac{1}{2}} \exp\left(\frac{w_1^2+w_2^2-2\theta w_1 w_2}{2(1-\theta^2)} - \frac{w_1^2+w_2^2}{2}\right)$ where $w_1 = \Phi^{-1}(u_1)$ and $w_2 = \Phi^{-1}(u_2)$	$\theta \in (-1, 1)$

X_i. Equation (1) shows that the dependence structure is modeled by a copula function and that the marginal densities can have different distributions.

Bivariate Copula Functions In this paper we propose the use of two copula functions: Frank and Gaussian, for modeling two-dimensional copulas. Both of the copula proposed are symmetric, and they can model both positive and negative dependencies and their respective joint distributions have equally low and high values[1]. Table 1 shows the density functions of the two copula:

Where $u_1 = F_1(x_1)$ and $u_2 = F_2(x_2)$ are the transformed values of the features x_1 and x_2 through their marginal distribution functions. Φ^{-1} is the quantile function of a normal distribution. θ is the dependence parameter. The estimation of θ is done through the maximum likelihood method.

2.2 Probabilistic Classifier with Copula Functions

Starting from the the Naive Bayes classifier equation, we can add information of the association between features using copulas. The probabilistic classifier would be as shown in Equation (2):

$$P(A|x_1, ..., x_d) = \frac{c(F_1(x_1|A), ..., F_d(x_d|A))P(A) \times \prod_{i=1}^{d} f_i(x_i|A)}{f(x_1, ..., x_d)}, \qquad (2)$$

where A represent one of the classes, $x_1, x_2, ..., x_d$, are the values of the features, and c is the joint density copula function.

3 Case Study

The data available for this paper comes from a collaboration with the Escuela Nacional de Estudios Superiores Unidad León, UNAM. Prior to their participation in the study, all participants provided informed consent after being thoroughly briefed on the study's protocol, which is non-invasive. Additionally, this research adheres to the ethical principles outlined in the Helsinki Declaration. A total of 58 records of individuals performing mandibular opening/closing, lateral and protrusion/retrusion movements were used. Of the 58 records, 29 come from individuals diagnosed with temporomandibular

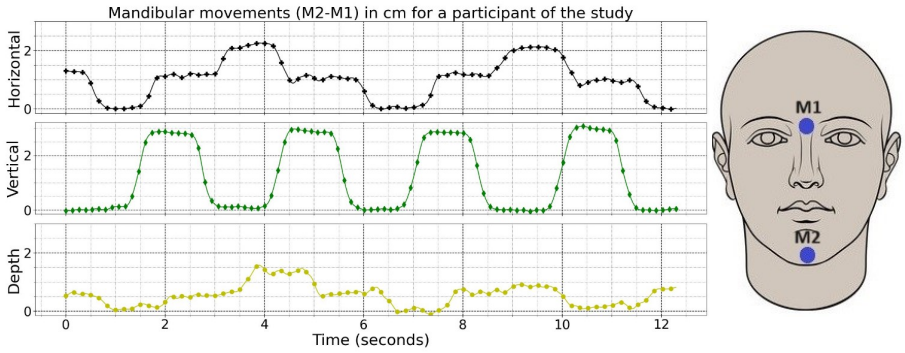


Fig. 1. Signals obtained from marker tracking.

arthralgia disorder and 29 records come from healthy individuals. For the clinical evaluation of the patient the DC-TMD[13] was used.

The records were obtained by tracking markers on the person's face. Two markers were placed in each person. Marker 1 was placed at the nasion craniometric point (between the eyes). Marker 2 was placed at the gnathion craniometric point (the chin). Marker 2 will record mandibular movement and marker 1 is a control marker for monitoring head movement. The equipment used for recording is an Intel Realsense Depth Camera D435. Tracking of the markers was carried out using the *OpenCV* library in Python.

The output of the tracking algorithms is a record of three signals per marker, corresponding to the three-dimensional position of the markers during mandibular movements. That is, each of the signals, shows us how the position of the marker changes with respect to the vertical, horizontal or depth axis over time. Figure 1 on the left shows segments of the signals obtained by tracking marker 2 for an individual. On the right it shows the location of marker 1 (M1) and marker 2 (M2) on the face as they were placed on the participants of the study.

4 Methodology

4.1 Preprocessing and Feature Extraction

The preprocessing consist of several stages: Initially a correction is applied to the signal to attenuate head movement artifacts from the mandibular movements. This is achieved by substracting the marker 1 signals from marker 2. Following this correction, the raw data from each recorded individual undergoes filtering using a 1D Gaussian filter and baseline correction. Subsequently the signals are segmented into three parts corresponding to the three mandibular movements recorded: Opening/closing (OC), Lateral (LAT) and Protrusion/Retrusion (PROT).

Table 2. Extracted features employed for TMD classification.

Biomechanical Features		
Name	Description	
OCY,OCX, OCZ	Max. deviation in y, x or z axis direction in opening movements	
LRX, LLX,LY,LZ	Max. deviation in x, y or z axis direction in lateral movements	
	Deviation in x axis measured to the left (LLX) and right (LRX)	
PY, PX, PZ	Maximum y, x or z axis deviation in Protrusion movements	
Speed, Acceleration and Jerk Features		
Name	Description	Formula
$\dot{s}_{max}$	Max speed	$\frac{\Delta s}{\Delta t}$
$\dot{s}_{avg}$	Mean speed	$\frac{\sum \dot{s}}{\lvert\dot{s}\rvert}$
$\sigma(\dot{s})$	Speed standard deviation	$\sqrt{\frac{\sum(\dot{s}-\dot{s}_{avg})^2}{\lvert\dot{s}\rvert}}$
$\ddot{s}_{max}$	Max acceleration	$\frac{\Delta \dot{s}}{\Delta t}$
$\ddot{s}_{avg}$	Mean Acceleration	$\frac{\sum \ddot{s}}{\lvert\ddot{s}\rvert}$
$\sigma(\ddot{s})$	Acceleration standard deviation	$\sqrt{\frac{\sum(\ddot{s}-\ddot{s}_{avg})^2}{\lvert\ddot{s}\rvert}}$
$\dddot{s}_{max}$	Max jerk	$\frac{\Delta \ddot{s}}{\Delta t}$
$\dddot{s}_{avg}$	Mean jerk	$\frac{\sum \dddot{s}}{\lvert\dddot{s}\rvert}$
$\sigma(\dddot{s})$	Jerk standard deviation	$\sqrt{\frac{\sum(\dddot{s}-\dddot{s}_{avg})^2}{\lvert\dddot{s}\rvert}}$
Frecuency Features		
Name	Description	Formula
P_{tot}	Total power	$\sum P_i$
P_{avg}	Mean power	$\frac{\sum P_i}{M}$
F_{ratio}	Frequency ratio	$\frac{\sum_{fs/2}^{fs} P_i}{\sum_{0}^{fs/2} P_i}$

When examining temporomandibular movements, various authors have proposed different feature extraction methods, including measurements such as maximum mouth opening [2,7], speed [3], Fourier transform [15], among others. In this study a collection of features previously proposed in other works for addressing TMD is used. Table 2 provides a summary of the ones extracted from the available data.

The speed, acceleration, jerk and frequency features were extracted for each of the signal segments corresponding to each of the mandibular movements. That is, the 12 proposed speed, acceleration, jerk and frequency features are extracted for each of the three signals (x, y and z axis) and for each of the movements (OC, LAT and PROT). That sums up to 108 features, to which we added the ten biomechanical features giving a total of 118 features.

4.2 Feature Selection

A subgroup of features was selected with the objective of dimensionality reduction. In this case the data was divided by the labels, having two groups:

Arthralgia group (AG) and the healthy group (HG). For each of the features extracted, a Wilcoxon test was applied to the groups. The Wilcoxon test compares the means of two groups, the p-value obtained by applying the test is a measure of how similar the means of the groups are. The lower the p-value, the larger is the difference in the means of the groups. From the original 118 features, the 20 features with the lowest p-values of the test were selected, Algorithm 1 shows the process in detail.

Algorithm 1 Feature Selection Pseudocode

```
Separate data in two subgroups AG and HG corresponding to each class.
for feature in datasets do
    Apply Wilcoxon test for the subgroups.
    Save the test p-value
end for
Sort in ascending order all p-values from the tests.
Select the 20 features with lowest p-values.
```

4.3 Copula Classifiers

Bivariate copula functions can model dependencies between two variables, but not all possible pairs of features show strong dependencies, therefore identifying which ones are the most relevant is a challenge. To this end, for the copula classifiers a tree graphical model based on Kruskal's minimum spanning tree algorithm(MST)[9] is proposed. Given a matrix $M_{d\times d}$, where d represents the total number of features, (in this case $d = 20$), contains the maximum log-likelihood for the combination of the i-th and j-th variables, according to the following equation (3):

$$M_{i,j} = \ell(\theta|u_i, u_j) = \sum \log c(\theta|u_i, u_j). \tag{3}$$

Using the MST algorithm requires transforming the maximum likelihood values into costs, which can easily be done by multiplying them by -1. When applied, the MST algorithm will render a tree-shaped graphic model showing the strongest dependencies to be modeled with a determined dataset. Algorithm 2 details the procedure for obtaining this graphical model.

Algorithm 2 is similar to the one proposed by [4] in the way the graphical model is obtained but with a few differences. One of them is that the criteria for determining the tree structure is log-likelihood instead of mutual information. Another is that the features in this work are continuous. Finally, the most important difference is that the dependence is modeled with copula functions. An example of the graphical model rendered by Algorithm 2 is shown in Figure 2. Then, Equation (4) defines a Bayes classifier based on bivariate copula functions and Figure 2:

Algorithm 2 Pseudocode for obtaining the tree graphical model

1: Separate training data by class and for each class dataset:
2: Estimate the distribution functions $F_1, F_2, ..., F_d$ of the features, for example, by using gaussian kernels
3: Transform the features $x_1, x_2, ..., x_d$ to $u_1, u_2, ..., u_d$ via $u_i = F_i(x_i)$
4: **for** i `in` $\{1, 2, .., d-1\}$ **do**
5: **for** j `in` $\{i+1, i+2, .., d\}$ **do**
6: Estimate $\theta_{i,j}$ parameter via maximum likelihood and save $M_{ij} = \ell(\theta|u_i, u_j)$
7: **end for**
8: **end for**
9: Multiply the matrix M by -1
10: Feed the matrix M to the MST algorithm

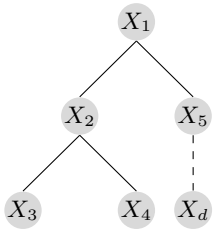


Fig. 2. Example of a dependence tree for a set of five features.

$$P(A|x_1, ..., x_d) = \frac{c(u_1, u_2)c(u_1, u_5)...c(u_{d-1}, u_d)P(A) \times \prod_{i=1}^{d} f_i(x_i|A)}{f(x_1, ..., x_d)}, \quad (4)$$

where $u_i = F_i(x_i)$ for $i = 1, 2, ..., d$. Then, going back to Figure 2, each of the edges connecting a pair of features represents a copula modeling the association between them. For the implementation of the copula classifiers the R package *MLCOPULA*[10] was used.

4.4 Model Training and Evaluation

A total of five models were trained for the classification task. Adding to the two copula models mentioned before (Frank and Gaussian), a Naive Bayes, KNN and Support Vector Machine were also trained. Each classifier is evaluated in randomly sampled partition cross validation scheme with 80% of the instances for training set and 20% for the validation set. Since the partitions are selected randomly, to ensure all instances are used in both training and validation, a high number of partition was proposed (50 partitions).

Four metrics were obtained for performance evaluation: accuracy, sensitivity, specificity and area under the ROC curve (AUC). The accuracy reflects the proportion of correctly classified instances, from both positive and negative classes; sensitivity shows the proportion of positive instances correctly classified,

Table 3. The 20 features selected via Algorithm 1 applied to the Features in Table 2.

OCX	OCY	OCZ	LRX
PZ	$OC\_Y\_\ddot{s}_{max}$	$OC\_Z\_\ddot{s}_{max}$	$OC\_Z\_\dddot{s}_{avg}$
$OC\_Y\_\dot{s}_{max}$	$OC\_Z\_\dot{s}_{max}$	$OC\_X\_\dot{s}_{avg}$	$OC\_Y\_P_{tot}$
$OC\_Y\_P_{ratio}$	$PROT\_Z\_\ddot{s}_{avg}$	$PROT\_Y\_\sigma(\dot{s})$	$PROT\_Z\_\dot{s}_{max}$
$PROT\_Y\_\dot{s}_{avg}$	$PROT\_Z\_\dot{s}_{avg}$	$PROT\_Z\_P_{avg}$	$LAT\_Y\_P_{tot}$

Table 4. Performance metrics by classifier. Best result in bold, standard deviation in parenthesis.

	Accuracy		Sensitivity		Specificity		AUC	
Classifier	Mean		Mean		Mean		Mean	
Naive Bayes	0.772	(0.12)	0.743	(0.20)	0.800	(0.17)	**0.833**	(0.13)
KNN	0.768	(0.12)	**0.800**	(0.16)	0.737	(0.17)	0.817	(0.12)
SVM	**0.782**	(0.10)	0.733	(0.15)	**0.830**	(0.16)	0.826	(0.10)
Frank	0.760	(0.11)	0.760	(0.21)	0.76	(0.17)	0.789	(0.14)
Gaussian	0.753	(0.11)	0.743	(0.20)	0.763	(0.18)	0.761	(0.14)

and specificity is the proportion of negative instances correctly classified. The AUC reflects the model's ability to distinguish between classes, the closer it is to 1, the better.

5 Results

The selected features using the method described in section 4 are shown in Table 3. The description of the features OCX, OCY, OCZ, LRX, PZ is shown in Table 2. To the rest of the features, the code OC, LAT and PROT represent the mandibular movement from which the feature was extracted: opening/closing, lateral and protrusion, respectively. The letter X, Y and Z represent the axis of the movement from which the feature was extracted: horizontal, vertical and depth, respectively.

Table 4 shows the mean and standard deviation of the measured metrics for each of the five classifiers tested. In bold is marked the best result for each of the metrics and in parenthesis the standard deviation. At first glance it is possible to see that although the performance of the copula classifiers is good overall, the SVM Classifier shows the best results in accuracy and specificity; the KNN had the best result for specificity and the Naive Bayes had the best AUC results.

Figure 3 shows the distribution of the accuracy and AUC for each of the classifiers. In terms of accuracy, the classifier with lower variance is the SVM, and is also the only one with atypical data points. The Gausssian copula, on the other hand, shows the highest variance of all the classifiers. Regarding the AUC, the SVM also shows the lowest variance and the Gaussian Copula the highest.

Figure 4 shows the distribution of the specificity and sensitivity for each of the classifiers. For the sensitivity metric, the boxplots show that the SVM

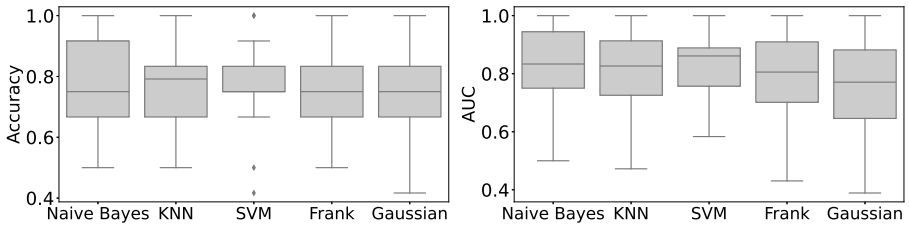


Fig. 3. Accuracy and AUC results by classifier.

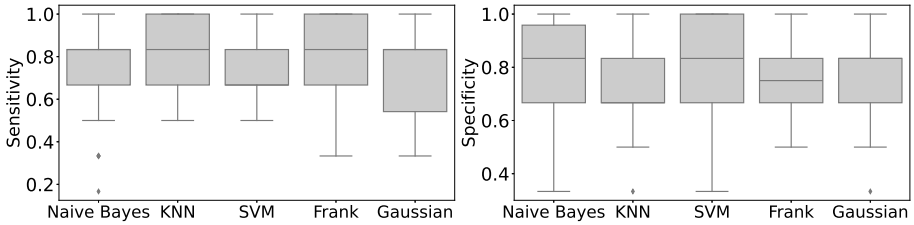


Fig. 4. Specificity and sensitivity results by classifier.

and Naive Bayes have the lowest variance, while the copula classifiers have the highest. The specificity boxplots, on the other hand, show the copula classifiers have the lowest variance.

To prove whether or not there is a difference in perfomance a Kruskal-Wallis test was applied to the results. Table 5 shows the p-values for the tests applied to the performance metrics. With a significance of $\alpha = 0.05$ the test shows that there are no differences in the means of the classifiers in the four metrics obtained.

6 Conclusions

This paper introduced copula classifiers with a tree graphical model for the classification of temporomandibular disorders. These classifiers were trained on movement data from healthy and TMD diagnosed individuals. Results show good mean performance in accuracy, sensitivity, specificity and AUC metrics for the copula classifiers. A comparison with other common classifiers used in similar tasks showed there is no statistical difference in performance for all the metrics used. Results also show that the copula classifiers have a higher variance in three out of the four metrics: accuracy, sensitivity and AUC.

Although the copula classifier performance is good, it was not particularly notable. This outcome may be related to the features selected for the training of the classifiers. Copula classifier perform best when there are strong relations

Table 5. Kruskal-Wallis test results for the performance metrics.

Metric	P-value	Metric	P-value
Accuracy	0.68130	AUC	0.07894
Sensitivity	0.3983	Specificity	0.05463

between the features in the data. Selecting the strongest dependencies among features is relevant in this proposed classifier because it could help differentiate between classes based on dependencies structures present in one class, but not in the other.

Since the selection method does not consider the strength of association among features, there is no guarantee that the selected features showed strong pairwise relationships. Another explanation for the performance is that the selected copulas do not provide the best model for the dependencies present in the data. Future work regarding copula classifiers for the prediction of temporomandibular disorders should consider a feature selection method accounting for feature dependencies. A set of copula functions modeling a wider range of dependencies should also be considered.

Acknowledgments. The first author Carlos López-Hernández acknowledges the financial support from the National Council of Humanities, Science and Technology of Mexico (CONAHCYT) given through the graduate scholarship program. The authors would like to thank Ulises Martín Arbaiza Martínez for applying the clinical evaluation of the participants in this study.

References

1. Balakrishna, N., Lai, C.D.: Continuous bivariate distributions. Springer Dordrecht Heidelberg London New York (2009)
2. Calil, B.C., Cunha, D.V.D., Vieira, M.F., Andrade, A.D.O., Furtado, D.A., Junior, D.P.B., Pereira, A.A.: Identification of arthropathy and myopathy of the temporomandibular syndrome by biomechanical facial features. BioMedical Engineering Online 19 (4 2020)
3. Castillo-Juárez, M.: Análisis cuantitativo de movimientos mandibulares para la clasificación de trastornos temporomandibulares por medio de un sistema de visión computacional. Tésis de Maestría, Universidad Autonoma de Aguascalientes (2022)
4. Chow, C., Liu, C.: Approximating discrete probability distributions with dependence trees. IEEE Transactions on Information Theory 14(3), 462–467 (1968)
5. Dowd, K.: Copulas in macroeconomics. Journal of International and global Economic Studies 1, 1–26 (7 2008)
6. Durham, J., Newton-John, T.R., Zakrzewska, J.M.: Temporomandibular disorders. BMJ (Online) 350 (3 2015)
7. Furtado, D.A., Pereira, A.A., Andrade, A.D.O., Peres, D., Junior, B., Silva, M.R.D.: A specialized motion capture system for real-time analysis of mandibular movements using infrared cameras (2013)

8. Genest, C., Favre, A.C.: Everything you always wanted to know about copula modeling but were afraid to ask. Journal of Hydrologic Engineering 12(4), 347–368 (2007)
9. Kruskal, J.B.: On the shortest spanning subtree of a graph and the traveling salesman problem. Proc. Amer. Math. Soc. 7, 48–50 (1956)
10. Salinas-Gutierres, R., Hernandez-Quintero, A., Montoya Calzada, A., Lopez Hernandez, C., Márquez Romero, J.: Mlcopula (2024), `https://cran.r-project.org/web/packages/MLCOPULA/`
11. Salinas-Gutiérrez, R., Hernández-Aguirre, A., Rivera-Meraz, M.J.J., Villa-Diharce, E.R.: Using Gaussian Copulas in Supervised Probabilistic Classification. Springer (2010)
12. Salinas-Gutiérrez, R., Hernández-Aguirre, A., Villa-Diharce, E.R.: Dependence trees with copula selection for continuous estimation of distribution algorithm. In: GECCO '11: Proceedings of the 13th annual conference on Genetic and evolutionary computation. pp. 585–592. ACM (2011)
13. Schiffman, E., Ohrbach, R., Truelove, E., Look, J., Anderson, G.: Diagnostic criteria for temporomandibular disorders (dc/tmd) for clinical and research applications: Recommendations of the international rdc/tmd consortium network * and orofacial pain special interest group † hhs public access. J Oral Facial Pain Headache 28, 6–27 (2014)
14. Sklar, A.: Fonctions de répartition á n dimensions et leurs marges. Publications de l'Institut de Statistique de l'Universite de Paris 8, 229–231 (1959)
15. Taşkıran, U., Çunkaş, M.: A deep learning based decision support system for diagnosis of temporomandibular joint disorder. Applied Acoustics 182 (11 2021)
16. Troka, M., Wojnicz, W., Szepietowska, K., Podlasiński, M., Walerzak, S., Walerzak, K., Lubowiecka, I.: Towards classification of patients based on surface emg data of temporomandibular joint muscles using self-organising maps. Biomedical Signal Processing and Control 72, 103322 (2 2022)
17. Zhang, Y., Wang, X., Liu, D., Li, C., Liu, Q., Cai, Y., Yi, Y., Yang, Z.: Joint probability-based classifier based on vine copula method for land use classification of multispectral remote sensing data. Earth Science Informatics 13, 1079–1092 (2020)

ISSN 1870-4069

Towards Semantic Data Extraction based on Long-Texts Clustering Using Parallel Computing

Diego Martínez-Maqueda[1], Cecilia Reyes-Peña[3], Jesús García-Ramírez[1,2]

[1] Universidad Politécnica Metropolitana de Hidalgo, Mexico

[2] Unidad Profesional Interdisciplinaria de Ingeniería Campus Tlaxcala, Mexico

[3] Instituto de Investigaciones en Matemáticas y en Sistemas, Mexico

231220009@upmh.edu.mx, ceciliareyes@turing.iimas.unam.mx, jegarciara@ipn.mx

Abstract. The World Wide Web and the development of increasingly complex websites and platforms generate large amounts of written data on a daily basis, which contain potential information that can be used for various purposes. The use of literary texts as a source of information for different Artificial Intelligence (AI) techniques is more limited compared to other types of texts, however, working with them can result in obtaining useful information as well as problems with their processing due to the peculiarities of these texts. This work proposes data pre-processing algorithms applied to short stories and novels written in Spanish for the subsequent extraction of information using pre-trained language models and other Natural Language Processing (NLP) techniques. In addition, the use of parallelization techniques is also included to optimize the execution times of some of the algorithms used.

Keywords: Text clustering, semantic similarity, natural language processing, parallel computing.

1 Introduction

The Big Data era is characterized by the massive access to diverse data types, including text, images, audio, video, and raw data. This abundance of web-derived data necessitates efficient processing techniques, as parallel computing, especially for machine learning applications. Incorporating metadata into the data analysis process can significantly enhance data comprehension, particularly for literary texts. Automatic information extraction techniques offer a promising solution for metadata acquisition.

Specialized literature encompasses a wide range of text sources, including medical notes, news articles, and even web resources like Wikipedia. However, the use of literary or fictional texts for information extraction remains limited.

This is because the context of literary texts can vary significantly, even within the same genre. Labaut and Bost [5] argue that analyzing fictional texts requires special attention, as improper techniques can lead to erroneous results. One approach to mitigate errors is to propose a genre-specific information extraction analysis. However, this necessitates the identification of more representative features, making the process more complex.

This work presents the implementation of a clustering algorithm specifically designed for Spanish literary texts (focusing on short stories and novellas). The algorithm aims to identify the most semantically similar texts based on extracted features. This facilitates the identification of named entities and the extraction of semantic relationships, ultimately leading to a comprehensive character description within each text. Additionally, we prioritize reducing information pre-processing runtime by leveraging parallel processing techniques to maximize computational efficiency.

The proposed clustering method successfully groups texts based on shared characteristics, such as author or country of origin. These well-defined clusters demonstrably enhance the effectiveness of information extraction tasks. Furthermore, the implementation of parallel computing techniques significantly reduces the computational time required for various tasks, including data vectorization and the elbow method.

This paper is structured as follows. Section 2 reviews existing research on named entity extraction and text vectorization. Section 3 details the proposed text clustering methodology, including the pre-processing steps. Section 4 presents the experimental findings. Finally, Section 5 discusses the conclusions and outlines potential future work.

2 Related Work

Named entity extraction is a relevant subtask within Natural Language Processing. NER aims to identify and classify specific words or phrases (entities) within a text according to predefined categories. These categories, often referred to as typological tags, can include people, organizations, locations, dates, monetary values, among others. In the next paragraphs we describe some related work to this task.

Van Dalen-Oskam et al. [3] adapted a system named namespace with conditional random fields (CRFs), support vector machines, and distributional word vectors. Their system achieved acceptable performance in identifying various entities (names, places, and organizations) within modern Dutch literary texts. The F1-measure results were 83.8%, 84.5%, and 89.3% for names, places, and organizations, respectively. Long et al. [7] focused on information extraction from complex materials like Qing and Ming dynasty novels. Their work employed CRFs to address the challenges posed by this genre. They achieved an F1-measure of 80.31%.

Bick [2] leverage a combination of named entity recognition (NER), the PALAVRAS morpho-syntactic and semantic analyzer [1], and an extension of

this method to identify various entities in Portuguese and Brazilian literature. Their focus extends beyond named entities, encompassing genres, titles, professions, social statuses, and familial relationships. While achieving good F1-measure results for character identification (63.4%) and genre identification (89.5%). Nonetheless, for the identification of professions (26.6% F1-measure) and familial relationships (15.5% F1-measure) yielded lower performance.

The aforementioned works in this section focus on named entity extraction but do not utilize parallel computing to expedite processing times. However, research by Fu et al. [4] demonstrates the potential of parallel computing in this domain. Their work employs tree-structured conditional random fields (CRFs) for named entity identification. The inherent parallelism of this approach, facilitated by the tree structure implementation, allows for significant reductions in processing times, as demonstrated by the authors.

If we talk about text clustering there are some researchers that work with literary texts. Omar [10] worked with a corpus of 74 novels applying a hybrid method to select the appropriate number of features to obtained good defined clusters. Using bag-of-words, he converted the texts into numeric representations obtaining a 74 x 37534 matrix. Then, with a variance analysis using ANOVA, a term frequency-inverse document Frequency (TF-IDF) analysis and a Principal Components Analysis, he reduced the number of features to 50. After that and with the new matrix, he used K-means method and obtained three good defined clusters.

Another research is by Sobchuk and Šeļa [11], they made combinations of level of thematic foregrounding, features analysis and measure of distance in literary texts clustering. The objective was to determine the best combination for the clusters. The results said that using a strong thematic foregrounding, which include lemmatization words, remove 100 most frequent words, remove entities, nouns, verbs, adjectives and adverbs and simplify the vocabulary replacing the less frequent words with their more frequent synonyms, the doc2vec algorithm to feature analysis and the cosine distance were the best combination.

On the other hand, Wang et al. [12] applied parallelization to the K-means algorithm to clustering texts. They divided all their process in three steps: first, they obtain a numerical representation of the words using the Wrod2Vec algorithm; second, they use the Canopy algorithm to calculate the initial centers and clusters; finally, they use de K-means algorithm to update the centers and obtain the final clusters using the Euclidian distance. In all this process, they applied parallel computing in the second and third step. They observed than increasing the number of processors, the acceleration ratio increased and the expansibility decreased.

3 Text Clustering through Parallel Computing

Extracting named entities effectively necessitates the development of methods that can capture descriptive features from the text and group them based on similarities. This task can be particularly challenging due to the

inherent complexities of literary text analysis. The diverse writing styles employed in literary works can further complicate the feature extraction and clustering processes.

To achieve this objective, we establish the following selection criteria for the text corpus used in the clustering process. We will select a dataset of short stories written in Spanish. Each story will be limited to a maximum of 15,000 words and must be categorized as either a short story or a novella. This focus on short narrative forms allows for a more controlled analysis while still enabling the exploration of potential feature variations between these categories. By employing these classifications (short story and novella), we aim to investigate the presence of distinct descriptive features within each class. We hypothesize that novellas, with their inherent narrative complexity compared to short stories, may exhibit a richer set of descriptive features. The documents were obtained from many different websites which offer a big variety of literary works. These texts were obtained manually and they were chosen based on their popularity. In general, the average length for all the texts is 17.84 words, but if we talk about novels the average length is 16.28 words and for the short stories the average length is 20.08 words.

To facilitate analysis and interpretation by a computational model, we propose the use of a numerical text representation that incorporates semantic information. We selected the Doc2Vec representation [6] for this purpose. Doc2Vec is a neural network-based approach that excels at generating numerical vector representations of text documents. This technique captures not only the surface meaning of the words themselves but also the broader context in which they appear within the document. This capability is particularly valuable for our task, as it allows us to account for the nuances of language use frequently encountered in literary texts. The decision of use doc2vec instead another option like sentence2vec is based on work with a only vector representation for each document that contains the particular context of them. In this point, we discarded the use of transformers because when working with long documents, the transformer could lose context information due to the concentration of attention on non-priority elements of the context.

After obtaining the numerical representation of the text data, we employ a dimensionality reduction technique called t-SNE [8] to visualize the high-dimensional vectors in a two-dimensional space. This visualization serves two purposes: 1) to assess the overall distribution of the data and 2) to guide the selection of an appropriate clustering method. The t-SNE shows a inverse symmetric data dispersion with a central axis parallel to dimension 2. Consequently, we propose using a distance-based clustering algorithm to identify the closest documents within each cluster based on their numerical vector representations.

Certain aspects of the proposed method are computationally expensive. To address this challenge, we leverage parallel processing techniques to accelerate the following tasks: text embedding transformation, data clustering, and hyperparameter optimization for the clustering algorithm. By employing

parallel computing, we anticipate a significant reduction in overall computation time compared to a sequential implementation. To achieve that we propose the use of programming language Python with the library multiprocessing which take advantages of devices with multiple processors. There are another languages that can work with parallel computing, for example C or C++ but, many projects related with machine learning and data science use Python as programming language. Moreno Arboleda et al. [9] made a comparative between Python and C++ in parallel algorithms and they demonstrated that for parallel process C++ has a better performance. For a small dataset, we expect that difference is no big.

Finally, we will evaluate the quality of the resulting clusters. For this purpose, we will employ various evaluation metrics, including the Silhouette coefficient. The Silhouette coefficient is a valuable metric that assesses the separation between clusters. It considers both the average distance between points within a cluster and the average distance to points in the closest neighboring cluster. A high Silhouette coefficient indicates well-separated clusters where points are closely grouped within their assigned cluster and far from points in other clusters.

4 Experimental Results

This section presents the experimental findings obtained from the proposed clustering method for short stories and novellas. The experiments address two primary objectives: (i) we aim to identify a clustering method that demonstrates effectiveness within this specific domain of literary text; and (ii) we evaluate the performance of our parallel implementation by testing it with datasets of varying sizes and utilizing different computational resources.

The experiments were conducted on a machine equipped with an Intel Core i5-9400F processor (6 cores) and 32 GB of memory. The operating system is Ubuntu 22.04. This configuration provided sufficient computational resources to effectively leverage parallel processing techniques.

Our experiments leverage a collection of 52 Spanish-language documents, comprised of 32 short stories and 20 short novellas. Associated metadata for each document includes title, year, author, and country of origin. Each document was converted into a numerical vector representation. Figure 1 depicts a t-SNE projection of this data in a two-dimensional space, which serves to visualize the distribution of the document embeddings.

Following the t-SNE dimensionality reduction, well-defined clusters may not be readily apparent in the data visualization (Figure 2). To address this challenge and determine the optimal number of clusters for the k-means algorithm, we employ the elbow method. This method involves running k-means with varying numbers of clusters and analyzing the resulting sum of squared errors (SSE) for each k value. The elbow method seeks the value of k where the SSE starts to plateau or decrease less significantly, indicating the point at which adding additional clusters yields diminishing returns in terms of intra-cluster variance reduction. The goal is to find a balance between minimizing intra-cluster variance

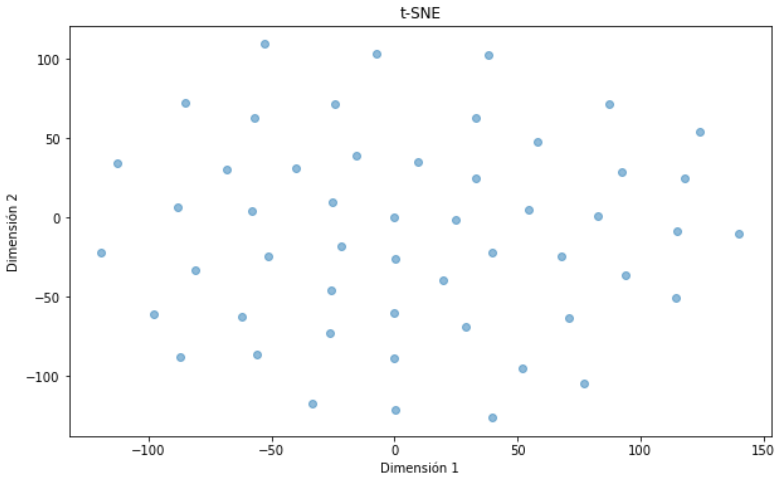


Fig. 1. Projection of the numerical data into a two-dimensional space using t-SNE algorithm.

(where points within a cluster are similar) and maximizing inter-cluster variance (where points between clusters are dissimilar). According to the results we consider that 3 is the best number of clusters.

The text transformation hyperparameters are described next:

- Vector Size: The dimensionality of the word vectors was set to 10, this value obtain good performance in our experiments.
- Minimum Count: This hyperparameter in Doc2Vec determines the minimum frequency a word needs to appear in the corpus to be included in the vocabulary. We set this value to 2.
- Epochs: The number of training epochs for Doc2Vec was set to 40.

The clustering results are visualized in the scatter plot of Figure 3, which utilizes t-SNE dimensionality reduction similar to Figure 1. The resume of meta data documents for each clusters is shown in Table 1, where there are the values of data and their frequencies within the clusters. The Silhouette coefficient score for this clustering is 0.16676, indicating a relatively low degree of separation between the clusters. This may be partially attributed to the presence of sparse documents within the dataset, which can introduce noise and hinder the clustering process.

In Table 1, the clusters obtained by the proposed method are shown. We can see that the embeddings capture information about the authors who wrote the novels or short stories. Similarly, the embeddings capture information about the years in which the texts were written. However, the embeddings do not perform well for the country or type of text (story or novella).

For the parallel implementation, we leverage the multi-process library in Python to facilitate multi-threading. This approach enables us to distribute tasks

Table 1. Summary of the obtained from the clustering pre-process, the numbers between the parentheses represent the frequency of the data obtained in each cluster.

	Cluster 0	**Cluster 1**	**Cluster 2**
Author	Herman Melville (1) Julio Cortazar (1) Gabriel García Marquez (2) Juan Rulfo (5) Amado Nervo (2) Albert Camus (1) Pedro Antonio de Alarcón (4) Jorge Luis Borges (3) Julia de Asensi (3) Frank Kafka (2) José Emilio Pacheco (1)	Horacio Quiroga (2) Edgar Allan Poe (2) Julio Cortazar (1) Juan Rulfo (1) Pedro Antonio de Alarcón (2) Julia de Asensi (1)	Ambrose Bierce (1) Juan Rulfo (4) Amado Nervo(2) Frank Kafka (1) Guy de Maupassant (1) Edgar Allan Poe (2) Gabriel García Márquez (1) Alberto Moravia (1) Fernando Díaz-Plaja (1) Pedro Antonio de Alarcón (1) José Emilio Pacheco (3)
Country	US (1) AR (4) CO (2) MX (8) FR (1) ES (7) GE (2)	UR (2) US (2) AR (1) MX (1) ES (3)	US (3) MX (9) GE (1) FR (1) CO (1) IT (1) ES (2)
Year	1853 (1) 1946 (1) 1961 (1) 1953 (5) 1968 (1) 1916 (1) 1917 (1) 1942 (2) 1877 (1) 1970 (1) 1941 (1) 1882 (1) 1972 (1) No date (7)	1917 (2) 1835 (1) 1964 (1) 1953 (1) 1843 (1) 1854 (1) No date (2)	2010 (1) 1953 (4) 1918 (1) 1915 (1) 1870 (1) 1832 (1) 1981 (1) 1957 (1)
Type	Story (14) Novella (11)	Story (6) Novella (3)	Story (13) Novella (5)

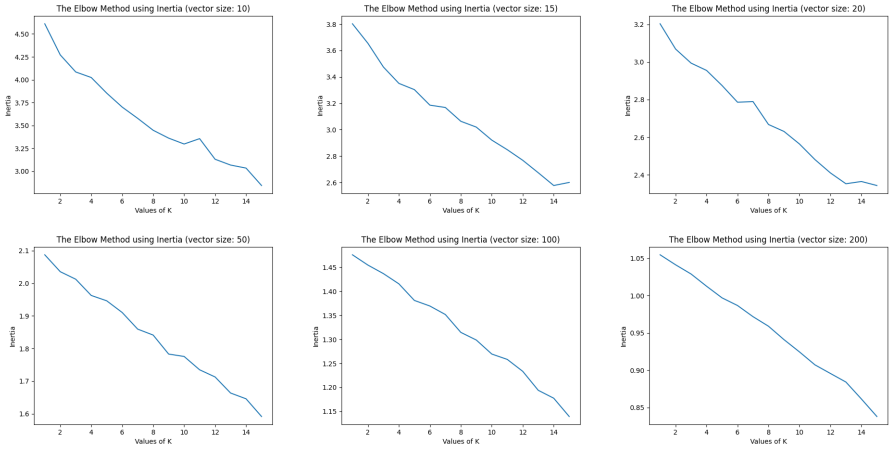


Fig. 2. Elbow method results using different vector size. According to the experimental results, the best value for the number of cluster is three.

across multiple cores, accelerating the computational processes. Additionally, we utilize NumPy for vectorization, which significantly optimizes operations involving numerical vectors.

In the first configuration, the text-to-numerical vector conversion and distance calculation tasks were parallelized, while the elbow method was executed sequentially. A reduction in execution time was observed primarily for the text vectorization step. This is likely because vectorization using libraries like NumPy is well-suited for parallelization. However, the overall speedup was limited because the elbow method remained sequential and potentially dominated the total execution time. In the Figure 4a depicts the execution time variations for the last configuration using different numbers of threads.

Building upon the observation that vectorization aided distance calculations in the first configuration, we designed a second configuration to explore potential speedups within the elbow method itself. This configuration leverages vectorization throughout the elbow method, aiming to reduce its computational cost. Our experiments demonstrated that this configuration achieved better overall performance compared to the first, as illustrated in Figure 4b.

Considering the execution time variations depicted in Figures 4, the second configuration demonstrably achieved superior performance compared to the first one. This highlights the potential benefits of incorporating vectorization within the elbow method itself. If the proposed method is to be implemented for larger datasets or more computationally intensive tasks, leveraging parallel processing techniques like those employed in the second configuration would be particularly advantageous to ensure efficient execution by maximizing the utilization of available computational resources.

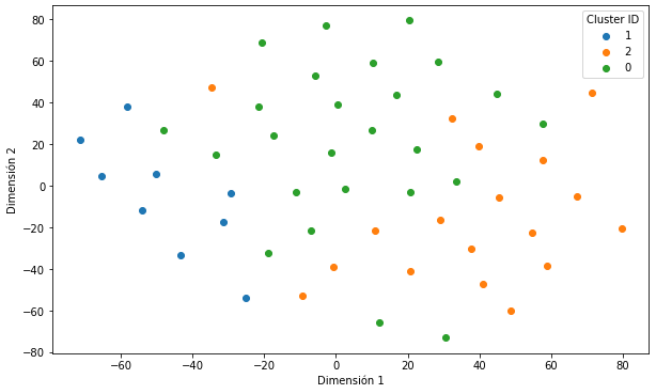
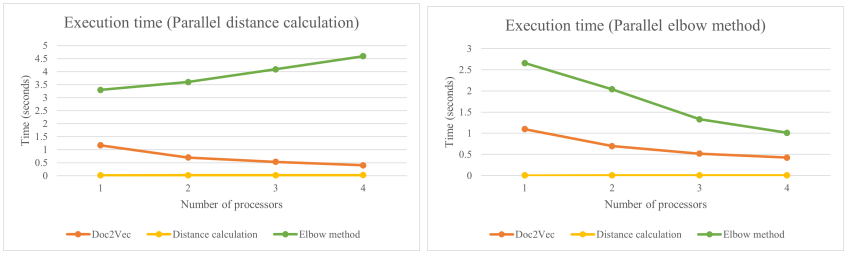


Fig. 3. Projection of the clusters obtained by the k-means algorithm.



(a) First Configuration (b) Second Configuration

Fig. 4. Execution time for the first (a) and second (b) configuration.

In addition to the execution time analysis, we can evaluate the effectiveness of parallel computing using two key metrics: speedup and efficiency. These metrics are calculated based on the following equations (Equations 1 and 2, respectively):

$$S_p = \frac{sequential\ time}{parallel\ time}, \tag{1}$$

$$Efficiency = \frac{S_p}{Number\ of\ cores}. \tag{2}$$

Tables 2 and 3 present the calculated speedup and efficiency values for the text vectorization and elbow method, respectively. The results indicate that for text transformation, the highest efficiency is achieved with two cores. In contrast, the elbow method demonstrates the best efficiency when using three cores.

These findings suggest that the optimal number of threads for parallel execution can vary depending on the specific task characteristics. The text vectorization process appears to benefit from a lower number of threads. Conversely, the elbow method might exhibit improved scalability with a slightly

Table 2. Speed up and efficiency for the experimental results for the numerical vector transformation.

Number of cores	Speedup	Efficiency
2	1.5797	0.7898
3	2.1161	0.7053
4	2.5917	0.6479

Table 3. Speed up and efficiency for the experimental results for the elbow method.

Number of cores	Speedup	Efficiency
2	1.3035	0.6517
3	1.9980	0.6660
4	2.6298	0.6574

higher number of threads, possibly because it involves more independent computations suitable for parallelization.

5 Conclusions

This work presented a pre-processing method for Spanish literary texts. The proposed method utilizes Doc2Vec to generate numerical representations of the texts, facilitating the identification of potential groupings through clustering techniques. We explored the effectiveness of parallel computing to improve the efficiency of key steps within the pre-processing pipeline, including text vectorization and distance calculations for clustering. The results demonstrate that parallelization can significantly reduce execution time, particularly for computationally intensive tasks like text vectorization.

The experimental results reveal an interesting interplay between vectorization and parallelization. For text vectorization, as observed in the figures, increasing the number of cores used generally leads to a significant improvement in execution time. This aligns with the strengths of vectorization libraries like NumPy, which are well-optimized for parallel processing.

In contrast, parallelization for distance calculation between points did not yield a substantial performance boost compared to vectorization alone. This suggests that for the current dataset size, the computational efficiency gained through vectorization the benefits of additional parallelization using multiple process. However, it is important to note that this behavior might change with larger corpora. As the number of texts in the corpus increases, the potential benefits of parallelization for distance calculations are likely to become more pronounced.

Future work will focus on extracting information from the pre-processed texts, such as character identification and feature extraction. Additionally, we plan to expand our experiments by incorporating a larger corpus of literary texts. This will allow us to gather more comprehensive data and validate the effectiveness of the proposed parallel processing configurations for a broader

domain with increased data volume. By analyzing larger datasets, we can gain a more robust understanding of how the method scales and identify potential optimizations for even more efficient text processing.

Acknowledgments. The authors thankfully acknowledge the computer resources, technical expertise and support provided by the Laboratorio Nacional de Supercómputo del Sureste de México, CONAHCYT member of the network of national laboratories, with the project 202301003N.

References

1. Bick, E.: PALAVRAS - A Constraint Grammar-Based Parsing System for Portuguese, pp. 279–302. Bloomsbury Academic (2014)
2. Bick, E.: Extraction of Literary Character Information in Portuguese: Extração de Informação sobre Personagens Literários em Português. Linguamática 15, 31–40 (2023)
3. van Dalen-Oskam, K., de Does, J., Marx, M., Sijaranamual, I., Depuydt, K., Verheij, B., Geirnaert, V.: Named Entity Recognition and Resolution for Literary Studies. Computational Linguistics in the Netherlands Journal 4, 121–136 (Dec 2014)
4. Fu, Y., Tan, C., Chen, M., Huang, S., Huang, F.: Nested Named Entity Recognition with Partially-Observed TreeCRFs. Proceedings of the AAAI Conference on Artificial Intelligence 35(14), 12839–12847 (May 2021), https://ojs.aaai.org/index.php/AAAI/article/view/17519
5. Labatut, V., Bost, X.: Extraction and Analysis of Fictional Character Networks: A Survey. ACM Computing Surveys 52(5), 1–40 (Sep 2020), https://dl.acm.org/doi/10.1145/3344548
6. Le, Q., Mikolov, T.: Distributed representations of sentences and documents. In: Xing, E.P., Jebara, T. (eds.) Proceedings of the 31st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 32, pp. 1188–1196. PMLR, Bejing, China (22–24 Jun 2014), https://proceedings.mlr.press/v32/le14.html
7. Long, Y., Xiong, D., Lu, Q., Li, M., Huang, C.R.: Named Entity Recognition for Chinese Novels in the Ming-Qing Dynasties. In: Dong, M., Lin, J., Tang, X. (eds.) Chinese Lexical Semantics, vol. 10085, pp. 362–375. Springer International Publishing, Cham (2016), series Title: Lecture Notes in Computer Science
8. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research 9(11) (2008)
9. Moreno Arboleda, F.J., Rincón Arias, M., Hernández Riveros, J.A.: Performance of Parallelism in Python and C++. IAENG International Journal of Computer Science, 50 (2023)
10. Omar, A.: Feature Selection in Text Clustering Applications of Literary Texts: A Hybrid of Term Weighting Methods. International Journal of Advanced Computer Science and Applications 11(2) (2020)
11. Sobchuk, O., Šeļa, A.: Computational thematics: comparing algorithms for clustering the genres of literary fiction. Humanities and Social Sciences Communications 11(1), 438 (Mar 2024), https://www.nature.com/articles/s41599-024-02933-6

12. Wang, H., Zhou, C., Li, L.: Design and Application of a Text Clustering Algorithm Based on Parallelized K-Means Clustering. Revue d'Intelligence Artificielle 33(6), 453–460 (Dec 2019), http://www.iieta.org/journals/ria/paper/10.18280/ria.330608

ISSN 1870-4069

Phonetic segmentation of the yuhmu language Using Mel-scale Spectral Representations

Eric Ramos-Aguilar[1,2], J. Arturo Olvera-López[1], Ivan Olmos-Pineda[1], Barbara Emma Sánchez-Rinza[3], Ricardo Ramos-Aguilar[2]

[1] Facultad de Ciencias de la Computación,
Benemérita Universidad Autónoma de Puebla,
Mexico

[2] Instituto Politécnico Nacional, UPIIT,
Mexico

[3] Facultad de Ciencias Físico Matemáticas,
Benemérita Universidad Autónoma de Puebla,
Mexico

eric.ramosag@alumno.buap.mx, {jose.olvera, ivan.olmos, barbara.sanchez}@correo.buap.mx, rramosa@ipn.mx

Abstract. The study of phonetic segmentation in indigenous languages of Mexico poses a challenge due to their linguistic and phonetic diversity. The use of digital signal processing techniques, machine learning, and both implicit and explicit segmentation, along with Mel-scale spectrogram analysis, provides an effective approach to identifying patterns that may delineate relevant information. Comparing the results with the actual number of phonemes in a word reveals both successes and areas for improvement. This paper proposes a methodology for segmental analysis in language Yuhmu considering parameter search in Mel scale, implementing cosine distance between spectrogram vectors, taking into account relevant data within the resulting matrices and seeking patterns of interest. The segment error rate yields results ranging from 23.89% to 26.03%, close to those reported in the literature on the subject.

Keywords: Phonetic segmentation, audio analysis, indigenous language.

1 Introduction

The analysis of language for automatic speech recognition, speaker verification, or language identification takes into account various forms of study, but always employing a segmentation of phrases or words, even using masking to conceal information and subsequently perform predictions and analysis. The aim is to find different semantic relationships that can be described through characteristic sets, some of which may be positional, frequency-based, word occurrence, or sets of words. It is possible to perform a segmental and suprasegmental analysis of texts, which considers sound, phoneme, articulation, phonology, intonation, and pauses [3].

Speech segmentation comprises two distinct processes, one of them manual and the other automatic (explicit and implicit). However, the central idea and research consideration involves conducting segmentation processes automatically, in order to provide more effective processes in the search for parameters of interest. The use of these techniques has served to enable different language analysts interested in phoneme search to interpret language and its constituent parts, considering lexical characteristics.

Phonetic segmentation of languages involves locating, sectioning, and delimiting the smallest part of a word, which has an interpretation in terms of phoneme, tone, and articulation.

The phoneme, being the smallest interpretive part of a word, considers elements of importance in the analysis because it is closely linked in the vast majority of cases with other phonemes that aid in the interpretation of a word, making it difficult to separate them. The phoneme can be described as the smallest unit of sound within a language system, when phoneme is changed for another within a context, the meaning of the word also changes [11].

Phonemes are conditioned by the meaning of a statement or by a family of sounds emitted in a particular language; therefore, any given expression can be transcribed using two phonetic levels of transcription: language-dependent (psychophonic) and language-independent (physiophonic)[12].

Phonemes tend to be interpreted differently depending on the language, considering tonalities that define a word. In some cases, these tonalities can be confused with intonations; however, tone is defined as contrasts with a paradigmatic dimension. The levels of contrast can be defined in some cases as high and low, although there may be contrasts of three, four, or five levels. Thus, a tonal language can be defined as a language that has a lexically significant, contrastive tone, but relative in each syllable [6].

In Mexico, according to INALI (Instituto Nacional de Lenguas Indígenas), there is a variety of linguistic families with different tonal interpretations, such as the Otomangue family, which includes languages like Mazahua and Otomí. Currently, the proposed research for voice segmentation and analysis of indigenous languages in Mexico focuses specifically on a purely manual analysis. This involves using software such as Praat or ELAN to analyze phonetic audio files, cutting words or sentences to identify phonemes and pronunciation tonalities [8].

In this paper, we propose to perform an explicit segmentation of phonemes in the Yuhmu language (Otomí from Ixtenco, Tlaxcala, Mexico). We will use words pronounced in the language that encompass all the phonemes, along with aligned phonetic representations in writing. This will be aided by Mel-scale spectrograms and pattern search using cosine distance between the information provided by the spectrogram matrix.

2 Related Work

Language analysis aided by neural networks involves processing a vast amount of data for training and validation, considering digital audio predominantly, with

transcriptions often incorporated as support. This entails aligning information between audio and text. When temporal alignment is conducted at the smallest unit of sound, the phoneme, it is referred to as speech segmentation [5].

Speech segmentation in indigenous languages of Mexico has been approached through manual analyses, utilizing digital audio representation, spectral analysis, and human perception. However, this process has been costly in terms of time and human resources. For example, [14] conducted manual segmentation using digital audio, speech signal envelope, spectrograms, and perception via Praat. Another study by [16] involved audio recordings of Náhuatl words pronounced by a single speaker, where ELAN software was used for subjective analysis of sound production, including segmentation for analysis purposes.

In [17], a description of intonational features of Hñöñhö (a variant of Otomí spoken in Tultepec, Querétaro, Mexico) is conducted through word segmentation using Praat. Praat and ELAN are considered open-source software for recording and analyzing words or phrases through: spectrograms, pitch, intensity, volume, and cochleograms using audio or video, respectively.

On the other hand, there are computational developments in automatic speech segmentation using two approaches: implicit and explicit [7]. Implicit segmentation involves supervised learning, where the system learns about the characteristics of each speech segment, then utilizes forced alignment of transcriptions using optimization. Explicit segmentation involves unsupervised learning; in some cases, it does not require forced alignment or utilizes pre-trained models for analysis processes, besides considering a reduced number of unlabeled data from pure audio [9].

An explicit segmental analysis is carried out in [9], where the boundaries of phonemes in a given utterance of Classical Arabic are identified using cosine distance similarity scores. This method achieves a total error rate of 14.48%, while the accuracy reaches 85.2% within a 10 ms alignment error.

Another segmental analysis is conducted using neural networks combined with a parameterized structured loss function, intending for the network to learn segmental representations to detect phoneme boundaries. This model utilizes the TIMIT and Buckeye corpora of English speech, achieving state-of-the-art performance in terms of F1-score and R-value, as mentioned in [10].

In [1], a phonetic segmentation analysis is conducted based on explicit (text-independent) segmentation using wavelet packet speech parametrization features and sparse representation classifier (SRC) utilizing Arabic and English (TIMIT), achieving a higher accuracy rate than the k-Nearest Neighbors (k-NN) on TIMIT.

Meanwhile, [4] proposes a single-scan method based on geometric quadrilaterals, which comprehends patterns of a speech signal. It utilizes the geometric nature of waveform trajectories, treating the input speech signal as a sequence of structural components. According to the paper, the algorithm's performance is assessed through experiments with spoken English words with an Indian native accent and Telugu sentences (an Indian language).

Phonetic segmentation is a significant problem involving the identification of boundaries of phonetic units. A challenge addressed by various researchers for over five decades. Most of the research focuses on the English language and languages with enough datasets for computational analysis. Methods like Hidden Markov Models and Deep Neural Networks require large amounts of data for training. However, there are current challenges in applying these approaches in the context of languages with limited resources [1].

This paper considers an explicit segmentation of phonemes in the Yuhmu language, which lacks a sufficient amount of data for deep learning. However, it is possible to automate segmentation processes and supervised learning. This analysis represents an initial praxis, as previously, phonetic analysis has been conducted manually through research by [2] using Praat. The intention of this process is to achieve automatic phoneme segmentation with the assistance of digital tools.

3 Yuhmu Language

Yuhmu is one of the variants of the Otomí language, spoken in Ixtenco, Tlaxcala, Mexico. It is endangered, as only a few elderly individuals (around 70 years old) maintain proficiency in its pronunciation. In some cases, individuals under the age of 60 understand the language, with no children learning it as their first language [2].

A community census by [2] indicates that there are currently around 100 speakers of the Yuhmu language, but their linguistic proficiency is not well understood. Additionally, Yuhmu lacks its own writing system, prompting efforts to represent its sounds phonetically or develop phonetic scripts.

Morphologically, Yuhmu comprises 32 phonemes represented in its pronunciations, as classified by the International Phonetic Alphabet (IPA), including 12 vowels (V) distributed between oral and nasal sounds, as shown in Fig. 1, where the expelled air exits through either the mouth or nose exclusively depending on the case. Additionally, it consists of 20 consonants (C) categorized by the location of the articulatory organs within the vocal tract. The airflow originates from the lungs, encountering obstruction from the lips or tongue. Unlike vowels, consonants may or may not be voiced, as depicted in Table 1.

For the analysis, 297 recordings of digital audio pronunciations of words were considered, encompassing all possible combinations of phonemes that form different words. Word structures in Yuhmu are generated from the following patterns: C-V, C-C-V, C-C-C-V, and C-V-V-V, these combinations may appear at the beginning, in the middle, and at the end of a word. In some cases, this structure may represent a single word. Additionally, the tone of the words is also considered, where it is possible to observe various words with high, low, and low-high tones [2].

The base dictionary used for the Yuhmu words is the one proposed by [2], which describes all the phonemes incorporated in the language and were analyzed subjectively. On the other hand, the digital recordings used have a

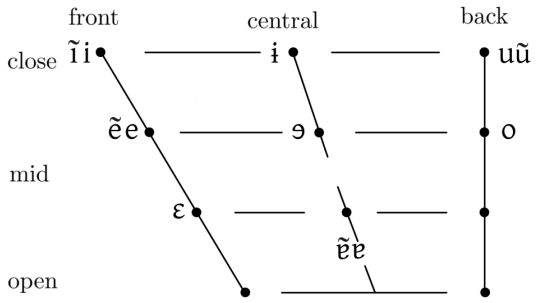


Fig. 1. Phonetic symbols of vowels (those with a tilde above them are considered nasal, while those without it are oral).

Table 1. Symbols of the International Phonetic Alphabet for Consonants in Yuhmu.

		Airway obstruction site					
Airway obstruction	Production of sound	Bilabial	Alveolar	Palatal	Velar		Glotal
Plosive	Voiceless	p	t		k	k^w	ʔ
	Voiced	b	d		g	g^w	
Affricates	Voiceless		ts	tʃ			
Fricative	Voiceless		s	ʃ			h
	Voiced		z				
Nasal	Voiced	m	n				
Tap or Flap	Voiced		ɾ				
Approximant	Voiced			j	w		

duration between 376 ms and 1.118 seconds, and they undergo preprocessing (noise attenuation, amplification, and word trimming).

4 Proposed Method

The following methodology is proposed to generate an explicit phonetic segmentation of the Yuhmu language, based on the analyzed literature, considering four phases (Fig. 2) described in the following sections.

4.1 Mel-scale Spectrograms

The search for optimal parameters is a fundamental aspect of research, as they contribute to improving the efficiency and effectiveness of phonetic segmentation. Therefore, we propose the parameters shown in Table 2, where spectral parameter search windowing is performed on the Mel scale. This is a technique

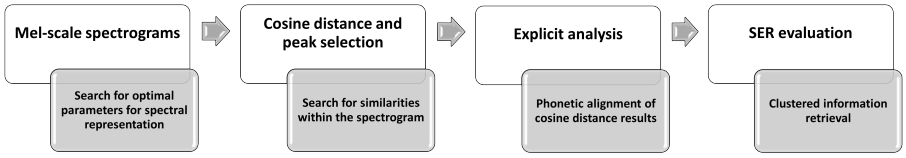


Fig. 2. Methodology phases for phonetic segmentation of Yuhmu.

Table 2. Selected parameters for optimal Mel-scale spectrogram search.

Window	Window size (ms)	Overlap (%)	Mel filter bands
Hanning	20, 25, 30, 35, and, 40	25, 50, and, 75	15, 20, 25, 30, 35, 40, and, 45

used in signal processing to analyze short segments of an audio signal. It involves dividing the audio signal into smaller segments, called windows, which are typically overlapped to capture temporal and frequency information more precisely. Each window undergoes a Fourier analysis to convert it from the time domain to the frequency domain, allowing visualization of the signal's energy at different frequencies in that time segment[13].

Mel scale is a frequency scale perceptually relevant to humans, and it is calculated through a non-linear transformation from frequency in Hertz to the Mel scale. The conversion of standard frequency to Mel frequency can be performed using equation 1, where F represents the frequency in Hertz (Hz) of a signal. For this research, a single Hanning window is considered, as it has been previously analyzed for this language in [15]:

$$F_{Mel} = 2595 Log_{10}(1 + \frac{F}{1000}). \quad (1)$$

On the other hand, the window size is proposed to vary from 20 to 40 ms with a 5 ms increment to ensure a reasonable interval for analysis. The overlap is set at 25%, 50%, and 75% of this window size to minimize information loss in each analysis during the process. Furthermore, although the literature suggests using 15 to 20 Mel-filter bands (Mels), which are filters used in audio signal processing to divide the frequency spectrum into bands that mimic human auditory perception, we propose an interval of 15 to 45 Mels to capture more information within the spectrogram and its resulting matrix. Since Mels represent the rows in the matrix, a higher number of Mels will result in a greater amount of information being represented.

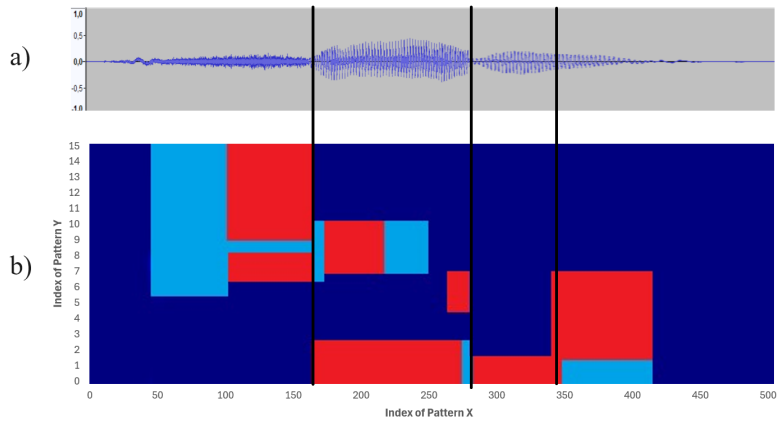


Fig. 3. a) Speech samples with phoneme boundaries over time. b) Hypothesis for ideal scores (Where navy blue tones represent null information, red tones represent high information, and light blue tones represent low information).

4.2 Cosine Distance and Peak Selection

The cosine distance (Eq. 2) is applied to the resulting spectrograms from the previous analysis. The purpose of this is to observe how similar the columns of the resulting spectrogram matrix are and to group information, separating the data corresponding to each of the phonemes. Ideally, we aim to establish a contrast of information (Fig. 3), considering the digital representation of audio and the segmentation generated from it. The cosine distance is defined as follows:

$$\text{cosine\_distance}(x, y) = \frac{\sum_{i=1}^{n}(x_i)^2 \cdot \sum_{i=1}^{n}(y_i)^2}{\sqrt{\sum_{i=1}^{n}(x_i)^2} \cdot \sqrt{\sum_{i=1}^{n}(y_i)^2} \cdot \sum_{i=1}^{n}(x_i \cdot y_i)}, \qquad (2)$$

where:

x_i and y_i are the components of vectors $\mathbf{x}$ and $\mathbf{y}$ respectively.

n is the length of the vectors (the number of components).

The symbol $\cdot$ represents the dot product between two vectors.

Following the cosine distance calculation between columns of the spectral matrix, a search for relevant information is conducted for each resulting matrix. Regardless of the parameter combination, the aim is to obtain the top 25% scoring points in each column, considering a threshold similar to the third quartile, which varies in each case, taking into account the number of Mels and windowing.

4.3 Explicit Analysis and SER Evaluation

Within the explicit analysis, a manual search for grouping information within the resulting matrices is considered. Segments found for each word are observed,

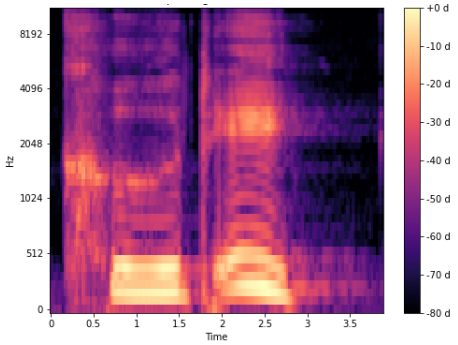


Fig. 4. Mel-scale spectrogram of the word 'ceñidor' in Yuhmu.

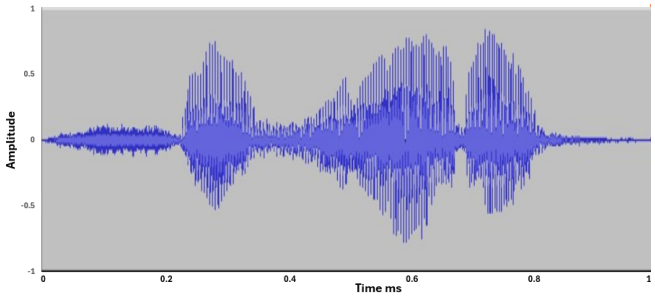


Fig. 5. Digital representation of an audio signal of the word "ceñidor" in Yuhmu.

the sample is representative, and delimited segments can be observed, recording the resulting number of segments in a table. Then, a comparison is made with the phonemes that each word actually contains. This comparison will be named Evaluation SER (Segment Error Rate), statistically indicating the segments generated by the proposed method and the percentage of error.

5 Experimental Results

When performing the methodological processes outlined in Section 4, we can obtain three matrices that are represented in figures, which describe the behavior of the information grouped in each of these. The main idea is to eliminate the connections that may link a phonetic representation to numerical information, considering three processes. The first one analyzes a Mel-scale spectrogram (see Fig. 4), which has as a precedent a digital audio signal represented in samples over time (see Fig. 5).

Having the representation in Mel-scale, cosine distance is applied between two vectors corresponding to the columns of the Mel-scale spectrogram matrix.

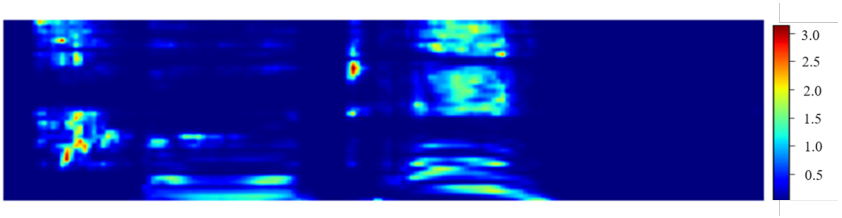


Fig. 6. Resultant matrix from cosine distance calculation of the word 'ceñidor' in Yuhmu.

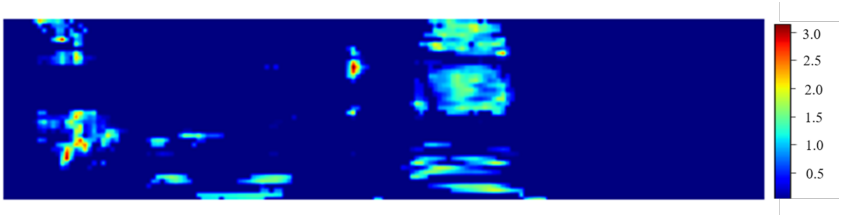


Fig. 7. Resultant matrix of maximum values per column of the word 'ceñidor' in Yuhmu.

These represent the spectral energy distribution of an audio segment based on the bands (for our case, the energy of the phonemes of the words in Yuhmu). Each column of the Mel-scale spectrogram shows how the spectral energy varies over time for a specific Mel frequency band. By calculating the cosine distance, scores are generated that provide a relationship between the values of each column. This will help us to create greater contrasts within the columns, generating high contrasts if they are dissimilar and low contrasts if they are similar, producing a separation in terms of energy and visualization of the spectrogram (see Fig. 6).

The process to obtain the search for contrasts still considers some connection between different groups of information representing the phoneme of a word. Thus, a selection of information from matrices is used, similar to the use of quartiles to obtain a threshold. However, in this case, it is obtained from the number of rows of the matrix, approximately 25% of high data for each column, which generates greater contrast between phonetic separation in the vast majority of words (see Fig. 7).

Finally, upon segmenting the data within the matrix, results of phonetic segmentation are obtained. Table 3 presents comparisons of the segmentation performed and the actual number of phonemes comprising a word. Ultimately, an average SER of 23.89% to 26.03% is obtained, varying among the combinations of windowing and the quantity of Mels used, which are represented in rows within the spectral representation (if the resulting spectral matrix was calculated with

Table 3. Comparison of some phonetic representations found in the spectrogram and actual phonemes, along with the resulting SER value.

Word 'yuhmu' represented in Spanish	Quantity of phonemes	20 ms (Windowing) 25 % (Overlap) 15 (Mels)	Difference between phonemes	SER %	20 ms (Windowing) 25 % (Overlap) 20 (Mels)	Difference between phonemes	SER %	20 ms (Windowing) 25 % (Overlap) 25 (Mels)	Difference between phonemes	SER %
abeja	4	4	0	0	4	0	0	4	0	0
abuela	3	2	1	33.33	2	1	33.33	2	1	33.33
abuelo	4	3	1	25	3	1	25	3	1	25
frijol	6	5	1	16.66	5	1	16.66	5	1	16.66
gallina	4	3	1	25	3	1	25	3	1	25
grueso	4	2	2	50	2	2	50	2	2	50
huazontle	6	4	2	33.33	4	2	33.33	4	2	33.33
niño	8	7	1	12.5	7	1	12.5	7	1	12.5
sal	1	1	0	0	1	1	0	1	1	0
viento	5	4	1	20	4	1	20	4	1	20

15 Mels, it will have 15 rows), providing energy information over time within the Mel scale.

The transformation from the Mel-scale spectrogram to the matrix selecting peaks considers groupings and contrast of information levels, achieving differences between the mentioned aspects regarding each phoneme. It is important to highlight that the results obtained from the average SER for each combination of spectral representation parameters of the words yield better results with a window size of 20 ms and a 25% overlap, along with a high number of Mels, ranging between 35-45. This consideration takes into account the methodological section, where a higher number of Mels implies a greater number of rows in the matrix, and therefore, more information in the final matrices. However, it does not represent a significant difference compared to the results with 15 to 35 filters.

Perhaps visually significant changes between Fig. 6 and Fig. 7 may not be perceived; however, there are datasets that are eliminated, generating a greater contrast between the datasets. At the matrix level, the given data are functional for performing groupings of these datasets, considering that we are not working at the image level but at the level of matrix information.

On the other hand, the SER considered for this experiment encompasses significant resultant values, as the literature cited displays segmentations with error rates ranging from 5-20%. However, these considerations involve languages with ample data for training and validating artificial neural networks. Taking this into account, the method presented yielded favorable results for a language with limited digital resources, encompassing all phonetic characteristics of the language.

6 Conclusions

Phonetic segmentation in indigenous languages of Mexico represents a significant challenge due to the linguistic and phonetic diversity of these languages. The results of our study indicate that the application of digital signal processing techniques, such as Mel-scale spectrogram analysis and cosine distance between

vectors, can provide an effective approach for phonetic segmentation in these languages. This is particularly relevant given the importance of preserving and adequately analyzing the phonetic structures of indigenous languages for linguistic purposes.

Comparing the results of phonetic segmentation with the actual number of phonemes in a word provides valuable information about the accuracy and effectiveness of the techniques used. This comparative analysis reveals both successes and potential areas for improvement in the segmentation process, which can guide future research efforts and methodological refinements in this field, ultimately proposing feedback on mispronunciation aided by its comparative and location of the mispronounced phoneme.

Phonetic segmentation in indigenous languages of Mexico is an evolving field that requires the application of interdisciplinary approaches. With the improvement of techniques and understanding of the unique phonetic characteristics of each indigenous language, it is possible to advance towards better documentation and preservation of these important cultural and linguistic expressions.

As a future work, it is planned to analyze a language with similar phonetic characteristics such as Jñatrjo (Mazahua from the State of Mexico), which belongs to the same linguistic family as Otomanguean languages and shares some similar words. However, Jñatrjo is distinct in having a developed system of writing and phonetic representation of words.

This segmental analysis aims to identify both good and poor pronunciation of words in both languages using a phonetic approach. This could significantly contribute to improving the understanding and preservation of the accurate phonetic structures of these indigenous languages, facilitating their proper documentation and cultural transmission.

Acknowledgments. This work is supported by CONAHCYT scholarship number 1169943, BUAP-VIEP 189 and UPIIT-IPN 20242590 research projects.

References

1. Al-Hassani, I., Al-Dakkak, O., Assami, A.: Phonetic segmentation using a wavelet-based speech cepstral features and sparse representation classifier. Journal of Telecommunications and Information Technology (4), 12–22 (2021)
2. Alarcon Montero, R.: Manual para la escritura de los sonidos del yuhmu. INAH (2023)
3. Ávila, J., Díaz, T., Ávila, C., Concepción, C., Carmona, E., Robaina, C., Cárdenas, C., Hernández, C., Mederos, P., Rouco, E.: Didáctica de la lengua española I. Editorial Pueblo y Educación (2013), https://books.google.com.mx/books?id=RR6owwEACAAJ
4. Bhagath, P., Das, P.K.: Quadrilaterals based phoneme segmentation technique for low resource spoken languages. In: TENCON 2022-2022 IEEE Region 10 Conference (TENCON). pp. 1–6. IEEE (2022)

5. Brognaux, S., Drugman, T.: Hmm-based speech segmentation: Improvements of fully automatic approaches. IEEE/ACM Transactions on Audio, Speech, and Language Processing 24(1), 5–15 (2015)
6. Gussenhoven, C.: The phonology of tone and intonation (2004)
7. van Hemert, J.P.: Automatic segmentation of speech. IEEE Transactions on Signal Processing 39(4), 1008–1012 (1991)
8. INALI: Catalogo de las lenguas indígenas nacionales. Retrieved in October 2022 from https://www.inali.gob.mx/clin-inali/ (Fecha de acceso: octubre de 2022)
9. Javed, M., Baig, M.M.A., Qazi, S.A.: Unsupervised phonetic segmentation of classical arabic speech using forward and inverse characteristics of the vocal tract. Arabian Journal for Science and Engineering 45, 1581–1597 (2020)
10. Kreuk, F., Sheena, Y., Keshet, J., Adi, Y.: Phoneme boundary detection using learnable segmental features. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8089–8093. IEEE (2020)
11. Kyriakopoulos, K.: Deep learning for automatic assessment and feedback of spoken english. Ph.D. thesis (2022)
12. Moore, R.K., Skidmore, L.: On the use/misuse of the term'phoneme'. INTERSPEECH (2019)
13. Pangaonkar, S., Panat, A.: A review of various techniques related to feature extraction and classification for speech signal analysis. In: ICDSMLA 2019: Proceedings of the 1st International Conference on Data Science, Machine Learning and Applications. pp. 534–549. Springer Singapore, Singapore (2020)
14. Penner, K.: Prosodic structure in ixtayutla mixtec: Evidence for the foot (2019)
15. Ramos-Aguilar, E., Olvera-López, J.A., Olmos-Pineda, I.: A general overview of language pronunciation analysis based on machine learning. Res Comput Sci 152 (2023)
16. Turnbull, R.: The phonetics and phonology of lexical prosody in san jerónimo acazulco otomi. Journal of the International Phonetic Association 47(3), 251–282 (2017)
17. Velásquez Upegui, E.P.: Entonación del español en contacto con el otomí de san ildefonso tultepec: enunciados declarativos e interrogativos absolutos. Anuario de letras. Lingüística y filología 8(2), 143–168 (2020)

ISSN 1870-4069

Dimensionality reduction based on Chi-Square Statistic and Testors for LGBT+phobia Detection

Metztli Ramírez-González, Jesús Ariel Carrasco-Ochoa,
José Francisco Martínez-Trinidad

Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE),
Mexico

{metztli.ramirez, ariel, fmartine}@inaoep.mx

Abstract. LGBT+phobia detection is a text classification task allowing identifying this kind of discrimination. However, text classification is a challenge due to the high dimensionality of word representations. In this work, we propose combining Chi-Square (χ^2) Statistic and irreducible testors to reduce dimensionality in LGBT+phobia Detection. The results in a public database (specially built for hate speech detection tasks) indicate that our proposal allows obtaining a small representation space to reach as good classification results as using the whole representation space's dimensionality for classifying and detecting LGBT+phobia in Mexican Spanish.

Keywords: LGBT+phobia detection, feature selection, testors.

1 Introduction

LGBT+phobia refers to any discrimination based on sexual preferences and/or gender identity [5], previously generalized as Homophobia [6]. According to data reported by the National Discrimination Survey (Enadis) 2022, more than 3.3 million people in Mexico report or have reported a non-normative sexual orientation or gender identity, representing approximately 3.6% of the national population. However, discrimination and violence towards the LGBT+ community do not occur in isolation; Enadis itself reports discrimination regarding access to housing, public office, health services, and even family rejection [7]. The CNDH [4] denounces that the LGBT+ community is often the victim of harassment, torture, arbitrary detention, and even murder, many times with total impunity.

According to the Mexican government [10], in the period from 2012 to 2022, Conapred has registered 1,175 complaints related to sexual and gender diversity. Moreover, according to Forbes [18], Mexico is the second country in Latam with the most hate crimes against the LGBT+ community. In Mexico, from 2019 to 2022, the LGBT+ community has been the victim of at least 305 violent acts motivated by hate, including murders, disappearances, and attempted murders, among others. Even so far, in 2024, at least 25 cases of murder against LGBT+ people have been reported [8].

These figures make this problem not only essential but also critical. These acts of discrimination find a space on social media, promoting and normalizing violence through hateful comments. Therefore, promptly detecting LGBTQ+phobic messages can improve the content moderation and create safer online environments [13].

LGBT+phobia can be included within the so-called hate speech. These messages are difficult to identify as they are influenced by various aspects such as the domain of a statement, its discursive context, concurrent media objects (images, videos, and audio), the historical and world context, and the identity of the author and recipient [21]. In recent years, research into hate speech detection has gained momentum, leading to exploring various techniques to address this problem.

Traditional methods typically leverage term frequency representations, which have shown promising results when combined with conventional classification approaches [23]. However, these representations produce high-dimensional spaces, which are difficult for classifiers to handle. Thus, the most recent solution approaches focus on the use of Transformers [9, 15–17, 20, 22, 26]; the problem with these approaches is that they often lack explainability.

To face the problems of explainability and high-dimensional representation spaces, in this work, this work proposes combining Chi-Square (χ^2) Statistic and irreducible testors [14] to reduce dimensionality in LGBT+phobia Detection. Instead of a transformer approach, we use the vocabulary of the problem domain and the frequency of terms in it. Our results in a public database show that our proposal can reduce the representation space while getting as good results as those by using the whole representation space's dimensionality for the classification and detection of LGBT+phobia in Mexican Spanish.

This work is based on the work developed in [19], where an alternative idea for addressing the identification of LGBT+phobia for the shared task HOMO-MEX 2024 for IberLEF [3, 12, 13] is proposed.

This paper is organized as follows: Section 2 summarizes the solutions given for the HOMO-MEX 2023 edition. Section 3 describes the proposed solution. Section 4 describes the obtained results. Finally, in section 5, we provide our conclusions and some directions for future work.

2 Related Work

HOMO-MEX is the first shared task focused on detecting LGBT+phobia in Mexican Spanish, organized for the first time in 2023 [1]. Mexican Spanish is characterized by its richness in language, such as the use of metaphors, allegories, figures of speech, insults, and nicknames that require a lot of context to be understood.

The proposals seen at Homo-Mex 2023 involved using Transformer-based models and data augmentation techniques [19]. Shahiki-Tash et al. [22] point out the importance of performing text preprocessing before using classification models so that they have better performance. Moriña et al. [17] used different

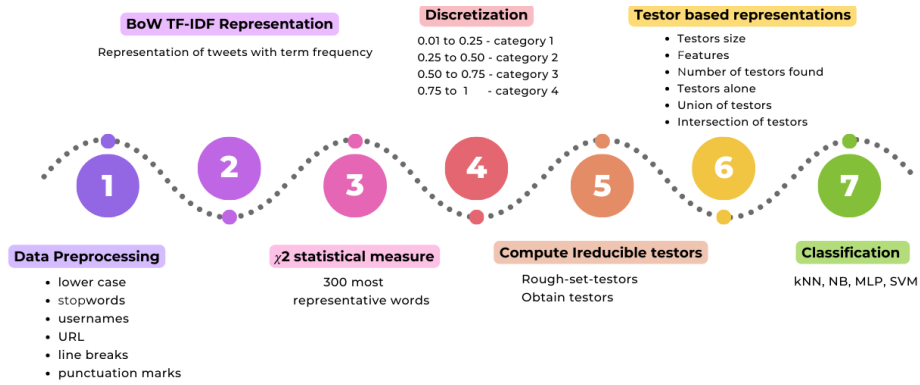


Fig. 1. Proposed solution diagram.

Transformer models and compared their performance. Marrugo-Tobón et al. [16] and Yigezu et al. [26] used data augmentation techniques and different variants of BERT for classification. García-Díaz et al. [9] experimented with combinations of Spanish and multilingual large language models (LLMs). Rosauro and Cuadros [20] compared traditional Classification Models and Transformers. Finally, Macias et al. [15] performed LGBT+phobia classification with different text representations based on frequency, without applying feature selection techniques and using classic models such as SVM and Bagging.

Most of these proposals focus on the use of Transformers and do not address feature selection, which is essential to characterize the problem domain. In this proposal, we focus on the frequency of the terms present in the problem vocabulary, applying dimensionality reduction based on χ^2 and irreducible testors to better understand the problem space.

3 Proposed Solution

Our approach is based on first applying a feature selection step over the BoW with TF-IDF representation using χ^2, and then a second feature selection step using irreducible testors. Figure 1 shows a schematic representation of the proposed approach, which consists of seven stages, we describe each of these stages in the following sections.

Data Preprocessing: All tweets are converted to lowercase and preprocessed by removing usernames, punctuation marks, URLs, line breaks, and stopwords.
BoW TF-IDF Representation: To represent the tweets, we used the classic Bag of Words (BoW) TF-IDF representation, which accounts for the term frequency in the document collection. Only those terms appearing in at least ten tweets were considered.
Chi-Square (χ^2) Statistic: From the BoW TF-IDF representation, the most relevant terms for class identification are calculated using the chi-square χ^2

statistic. The 300 terms with the highest statistical values are selected to represent the data.

Discretization: With the 300 selected terms, for allowing computing irreducible testors, a new BoW TF-IDF representation is built. Based on this representation, values are discretized into the following ranges: 0.01 to 0.25 - category 1, 0.25 to 0.5 - category 2, 0.5 to 0.75 - category 3, 0.75 and above - category 4.

Compute Ireducible testors: The RCC-MAS algorithm is applied on the discretized representation to find some irreducible testors. A testor is a subset of features that can distinguish objects from different classes in a data set, i.e., no object of one class is confused with any object of another class [14]. In our proposed approach, a testor becomes a set of words selected based on their frequency in the data. These words constitute a sufficient vocabulary to represent the data and allow us to differentiate between the classes. RCC-MAS is an algorithm designed to calculate all constructs; however, we employed it according to [11] to compute irreducible testors, but given that the search space for searching irreducible testors is very large, the algorithm is given a fixed execution time to just finding some testors.

Testor-based representations: The obtained irreducible testors are analyzed to determine their characteristics, such as the number of irreducible testors found, the words they contain, and their size. To evaluate the irreducible testors, the data representation is modified so that the BoW TF-IDF, previously reduced to 300 words, is characterized only by the words included in a testor, and then classification is performed. Additionally, representations and classifications are evaluated with the intersection and union of the identified irreducible testors. Finally, the results obtained using the 300 words are compared with those obtained using the words identified by the irreducible testors.

Classification: From each testor-based representation, the classification is performed employing k-Nearest Neighbors with k=10, Naive Bayes, Multilayer Perceptron, and SVM, using 5-fold cross-validation. The classification quality was measured using F1-score.

4 Experiments

4.1 Dataset

To evaluate the methodology we use the dataset provided by HOMO-MEX. HOMO-MEX is the first corpus for the detection of LGBT+ phobia in Mexican Spanish. It comprises public tweets extracted using the Twitter API, including keywords used in LGBT+phobic contexts. First, a list of nouns used to refer to the LGBT+ community was compiled (see Figure 1). Then, more than ten thousand tweets containing any of these nouns from the last ten years were selected. Four annotators subsequently labeled each tweet as LGBT+phobic, non-LGBT+phobic, or not related to the LGBT+ community [25]. HOMO-MEX hate speech detection task [3, 5, 12]: has a total of 8800 training data, divided into 5482 instances for the Non-LGBT+phobic class, 1072

Table 1. LGBT+phobia detection results on the different representations used in our experiments (F1-score).

		Classifiers			
Representation	**Size in words**	**SVM**	**kNN**	**NB**	**MLP**
BoW- All words	1191 words	**0.833332**	0.803367	0.775447	0.823568
300 words	300 words	0.819405	0.801503	0.681794	**0.823116**
Testor 1	133 words	0.803647	0.790423	0.458807	0.801504
Testor 2	133 words	0.803909	0.792035	0.460065	0.803985
Testor 3	134 words	0.805923	0.797587	0.460802	**0.811365**
Testor 4	134 words	0.805466	0.793303	0.450702	0.810714
Testor 5	134 words	0.802722	0.789211	0.455831	0.804951
Testor 6	134 words	0.802229	0.79703	0.452902	0.804168
Union	137 words	0.806558	0.798218	0.461325	0.803276
Intersection	131 words	0.802903	0.799502	0.459348	0.804686

instances for the LGBT+phobic class, and 2246 instances for the irrelevant class.

HOMO-MEX has several subtasks, but this work focuses on LGBT+phobic speech detection, which aims to predict the label of each tweet. It is a multiclass task in which a tweet can belong to one of the next three classes:

- LGBT+phobic (P), which includes tweets that contain hate speech directed against persons whose sexual orientation and/or gender identity differs from heterosexuality.
- Non-LGBT+phobic (NP), which includes tweets that mention concepts related to the LGBT+ population but without any intention of hate speech.
- Non-LGBT+ related tweets (NR) those that have no relationship with the LGBT+ community.

4.2 Application of the Proposed Solution

Six different testors were obtained by applying the proposed methodology and the RCC-MAS algorithm. Testors 1 and 2 contain 133 words each, while Testors 3, 4, 5, and 6 contain 134 words each. Combining these testors (union) results in 137 words, and the intersection results in 131 words. Table 1 shows the results of the experiments on LGBT+phobia detection (in terms of F1-score) using the entire vocabulary, the 300 most representative words, and the testor-based representations, including their union and intersection.

4.3 Analysis of Results

Some highlights found after analyzing the results of our experiments are:

Best Results: The best results were obtained using the entire vocabulary without attribute selection (1191 words). However, the results obtained with the initial choice of 300 words (F1-score of 0.823116) and the 134-word testors (F1-score of 0.811365) are not very different from the best result (F1-score of 0.8333). Applying the Kruskal-Wallis test, a p-value of 0.437 is obtained for all classifiers, indicating that there are no statistically significant differences between the different word representations. Furthermore, according to the t-test, there is no statistical difference when comparing the use of all the terms against the 300-word selection by chi-square statistic(χ^2) or against the testor-based representations. Additionally, the results of the Nemenyi test do not show any significant variation when comparing pairs of representations. Therefore, we can conclude that there are no significant differences between all the representations tested. This suggests that the 134 words selected by the irreducible testor three are sufficient to adequately represent the data and provide acceptable classification performance. It is also important to highlight that words contained in the six testors are almost the same, with a variation of only 6 words. Thus, we can conclude that the vocabulary sufficient to characterize the entire data set is about 131 words.

Selected Words: The 137 words chosen by at least one irreducible testor are listed below:
'persona', 'putas', 'nomás', 'transexual', 'si', 'mariconcito', 'acuerdo', 'maría', 'borracho', 'loquita', 'personas', 'estupido', 'bis', 'facewithrollingeyes', 'gays', 'comunidad', 'mujeres', 'jajajajaja', 'mariquita', 'acá', 'cis', 'hoy', 'feministas', 'maricas', 'loveislove', 'discriminación', 'paisanos', 'show', 'hombres', 'editorial', 'clóset', 'llorar', 'parejas', 'at', 'puto', 'lucha', 'gay', 'feminista', 'maricones', 'identidad', 'gente', 'lea', 'sinónimo', 'liosas', 'nombre', 'problema', 'sánchez', 'mismo', 'sexo', 'vida', 'maricon', 'género', 'pareja', 'súper', 'volviendo', 'sorprende', 'ser', 'vato', 'mariquitas', 'historia', 'queer', 'bi', 'in', 'veces', 'pues', 'chiva', 'campeonato', 'puro', 'lesbiana', 'puros', 'primera', 'aclaro', 'diciendo', 'putito', 'musicalnotes', 'hablar', 'negras', 'solo', 'pinche', 'santa', 'transformer', 'ke', 'hetero', 'homosexuales', 'vergas', 'respeto', 'tragedias', 'bisexual', 'bisexuales', 'loca', 'dragas', 'pinches', 'jotos', 'vestidas', 'heterosexuales', 'putos', 'madre', 'bandera', 'enseñando', 'joto', 'maricón', 'paso', 'méxico', 'hombre', 'chicos', 'lesbianas', 'travesti', 'tema', 'the', 'gayboy', 'transexuales', 'gol', 'aguantas', 'mujer', 'locas', 'drag', 'marica', 'odiantes', 'puta', 'raritos', 'vuelta', 'sexual', 'jotito', 'vestida', 'homosexual', 'transgénero', 'transformers', 'rarita', 'trans', 'derechos', 'puñal', 'jajaja', 'papel', 'draga', 'calor', 'puñetas'.

4.4 Significance of Words by Class

The most significant words according to the Chi-Square (χ^2) Statistic for each class are shown below:

– **Non-LGBT+phobic class:** trans, gay, gente, mujeres, homosexual, homosexuales, lesbiana, mujer, drag.

- **LGBT+phobic class:** joto, marica, jotos, maricon, puto, mariquita, maricas, gay, puñal, pinche.
- **Non-LGBT+ relate Class:** loca, vestida, puta, puto, bi, locas, transformers, hoy, bis, madre.

This analysis of the vocabulary used in each class shows marked differences, especially in the presence of insults and derogatory words in the LGBT+phobic class, in contrast to the other two classes that have fewer terms of this type. This language analysis can help identify and build a lexicon-based solution.

5 Conclusions

In this paper, we present a different approach to detecting LGBT+phobia in Mexican Spanish, by applying dimensionality reduction techniques, first a reduction to 300 vocabulary words selected using chi-square χ^2, and a further reduction using irreducible testors, obtaining a space of 134 words that seems sufficient to represent the whole dataset. LGBT+phobia detection results are reasonable compared to using other approaches, and this approach based on dimensionality reduction techniques has the property of knowing which specific words or terms we are representing the LGBT+ phobic tweets, which is lost if we apply the approach based on transformers, therefore, with the proposed approach in this paper, we contribute explainability to the problem.

We can conclude that linguistic analysis is a great tool for understanding the problems related to hate speech, and this approach can be beneficial and even improve some of the current techniques. In future work, we will evaluate the usefulness of constructs [24] and Goldman fuzzy reducts [2] from rough set theory, which can work directly on the BoW TF-IDF representation without discretizing it.

Acknowledgments. This research was partially supported by the National Council of Humanities, Sciences, Technologies, and Innovation of Mexico (CONAHCyT) through its graduate study scholarship program.

References

1. Bel-Enguix, G., Gómez-Adorno, H., Sierra, G., Vásquez, J., Andersen, S.T., Ojeda-Trueba, S.: Overview of homo-mex at iberlef 2023: Hate speech detection in online messages directed towards the mexican spanish-speaking lgbtq+ population. Natural Language Processing 71, 361–370 (2023)
2. Carrasco-Ochoa, J., Lazo-Cortés, M., Martínez-Trinidad, J.: An algorithm for computing goldman fuzzy reducts. In: Pattern Recognition: 9th Mexican Conference, MCPR 2017, Huatulco, Mexico, June 21-24, 2017, Proceedings. Lecture Notes in Computer Science, vol. 9, pp. 3–12. Springer, Cham (2017)
3. Chiruzzo, L., Jiménez-Zafra, S.M., Rangel, F.: Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages.

In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org (2024)
4. CNDH: Día internacional contra la homofobia, la transfobia y la bifobia. https://www.cndh.org.mx/noticia/dia-internacional-contra-la-homofobia-la-transfobia-y-la-bifobia-0, (accessed 01 July 2024)
5. CodaLab: HOMO-MEX: Hate speech detection towards the Mexican Spanish speaking LGBT+ population. https://codalab.lisn.upsaclay.fr/competitions/10019, (accessed 01 July 2024)
6. CONAPRED: Guía para la acción pública contra la homofobia. México, D. F. (2012)
7. CONAPRED: Discriminación en contra de las personas por su orientación sexual, características sexuales e identidad y expresión de género. México, D. F. (2024)
8. FundaciónArcoiris: Observatorio nacional de crímenes de odio contra personas lgbt. http://www.fundacionarcoiris.org.mx/agresiones/panel, (accessed 02 July 2024)
9. García-Díaz, J.A., Jiménez-Zafra, S.M., Valencia-García, R.: Umuteam at homo-mex 2023: Finetuning large language models integration for solving hate-speech detection in mexican spanish. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) (2023)
10. GobiernoMéxico: Registra conapred mil 175 quejas relacionadas con personas de la diversidad sexual y de género. https://www.gob.mx/segob/prensa/registra-conapred-mil-175-quejas-relacionadas-con-personas-de-la-diversidad-sexual-y-de-genero, (accessed 01 July 2024)
11. González-Díaz, Y.: Rcc-mas. GitHub. https://github.com/ygdiaz1202/RCC-MAS/blob/master/out/artifacts/RCC_MAS_jar/RCC-MAS.jar, (accessed 05 June 2024)
12. Gómez-Adorno, H., Bel-Enguix, G., Calvo, H., Vásquez, J., Andersen, S., Ojeda-Trueba, S., Alcántara, T., Soto, M., Macias, C.: Overview of homo-mex at iberlef 2024: Hate speech detection towards the mexican spanish speaking lgbt+ population. Natural Language Processing 73 (2024)
13. HOMO-MEX: Homo-mex 24: Hate speech detection towards the mexican spanish speaking lgbt+ population. https://sites.google.com/view/homomex/home, (accessed 01 July 2024)
14. Lazo-Cortes, M., Ruiz-Shulcloper, J., Alba-Cabrera, E.: An overview of the evolution of the concept of testor. Pattern Recognition 34(4), 753–762 (2001)
15. Macias, C., Soto, M., Alcántara, T., Calvo, H.: Impact of text preprocessing and feature selection on hate speech detection in online messages towards the lgbtq+ community in mexico. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) (2023)
16. Marrugo-Tobón, D.A., Martinez-Santos, J.C., Puertas, E.: Natural language content evaluation system for multiclass detection of hate speech in tweets using transformers. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) (2023)
17. Moriña, A.J.M., Pásaro, J.R., Vázquez, J.M., Álvarez, V.P.: I2c-uhu at iberlef-2023 homomex task: Ensembling transformers models to identify and classify hate messages towards the community lgbtq. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) (2023)

18. México, F.: Pese a los avances legales, méxico lidera en crímenes de odio contra personas lgbt. https://www.forbes.com.mx/mexico-lidera-crimenes-odio-personas-lgbt-avances-legales/, (accessed 02 July 2024)
19. Ramírez-González, M., Hernández-Farías, D., Montes-y Gómez, M.: Labtl-inaoe at homo-mex 2024: Distance-based representations for lgbt+ phobia detection. In: XL Congreso Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN 2024) (2024), (In press)
20. Rosauro, C.F., Cuadros, M.: Hate speech detection against the mexican spanish lgbtq+ community using bert-based transformers. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) (2023)
21. Schmidt, A., Wiegand, M.: A survey on hate speech detection using natural language processing. In: Proceedings of the fifth international workshop on natural language processing for social media. pp. 1–10 (2017)
22. Shahiki-Tash, M., Armenta-Segura, J., Ahani, Z., Kolesnikova, O., Sidorov, G., Gelbukh, A.: Lidoma at homomex2023@iberlef: Hate speech detection towards the mexican spanish-speaking lgbt+ population. the importance of preprocessing before using bert-based models. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) (2023)
23. Shridhara, M., Pristaš, V., Kotvytskiy, A., Antoni, L., Semanišin, G.: A short review on hate speech detection: challenges towards datasets and techniques. In: Proceedings of the 2023 World Symposium on Digital Intelligence for Systems and Machines (DISA). pp. 204–209. IEEE (2023)
24. Susmaga, R.: Reducts versus constructs: an experimental evaluation. Electronic Notes in Theoretical Computer Science 82(4), 239–250 (2003)
25. Vásquez, J., Andersen, S., Bel-Enguix, G., Gómez-Adorno, H., Ojeda-Trueba, S.L.: Homo-mex: A mexican spanish annotated corpus for lgbt+ phobia detection on twitter. In: Proceedings of the 7th Workshop on Online Abuse and Harms (WOAH). pp. 202–214 (2023)
26. Yigezu, M.G., Kolesnikova, O., Sidorov, G., Gelbukh, A.: Transformer-based hate speech detection for multi-class and multi-label classification. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023) (2023)

ISSN 1870-4069

Image Segmentation based on Division and Region Fusion

Samantha Acosta-Ruiz

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
Mexico

224570157@viep.com.mx

Abstract. This article explores digital image processing with a focus on fundamental segmentation techniques. It highlights two main approaches: region division and fusion, and border-based segmentation, each suited for specific analysis objectives. Region merging and border segmentation are key for identifying areas like housing and roads in satellite images. Quantitative evaluations were performed on metrics such as mean square error, Jaccard index (IOU), and DICE coefficient. The effectiveness of methods such as Sobel, Roberts, and Laplacian over the conventional Canny edge detector in the Hough transform was noted. Techniques such as the Otsu method for thresholding, noise removal using median and Gaussian filters, and edge detection were essential. Gabor filters were crucial for highlighting textures and distinguishing features like vegetation and empty areas. Adaptability and experimentation were key, with the strategic combination of techniques, including the optimized Hough transform, proving effective in accurately delineating populated areas and roads. The study underscores the importance of selecting and combining the right techniques for optimal image processing results.

Keywords: Region fusion, Gabor filters, Hough transform, segmentation.

1 Introduction

A satellite image is not a film-based photograph, almost all commercial remote sensing satellites capture images using digital sensors based on the same principles as digital cameras. A satellite sensor has thousands of tiny detectors that measure the amount of electromagnetic radiation reflecting the Earth's surface and the objects on it. These measurements are called spectral, and each spectral reflectance value is recorded as a digital number. These numbers are transmitted back to Earth where, through a processing, they are converted into colors or shades of gray to create an image that looks like a photograph [1].

An image can be defined as a two-dimensional function of light intensity $f(x, y)$, where x and y represent the spatial coordinates and the value of f at any point (x, y) is proportional to the brightness (or gray level) of the image at

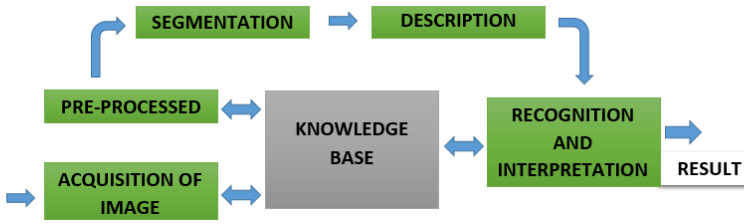


Fig. 1. Stages of image processing.

that point. A digital image is an $f(x, y)$ image that has been discretized both in terms of spatial coordinates and brightness; it can be considered as a matrix whose row and column indices identify a point of the image and the value of the corresponding element of the matrix indicates the gray level at that point. The elements of such a digital design are called picture elements or more commonly pixels, an abbreviation of their English denomination "picture elements" [5].

Figure 1 shows the necessary basic steps in image processing. The process starts with the image acquisition stage, where an image sensor produces signals that must be digitized. For example, light is used for photography; X-rays for radiography, ultrasound for sonography, etc. The nature of the sensor will depend on the type of application to study. The next stage is the preprocessing, carried out in order to detect and eliminate any flaws that may exist in the image to improve it. The most commonly used techniques at this stage are: a) contrast enhancement, b) noise removal, and c) restoration. In the following segmentation stage, the image is divided into its constituent parts or objects in order to separate the necessary processing parts from the rest of the image that are not of interest for the desired application. The basic techniques at this stage are those oriented to: a) the pixel, b) the edges, and c) the regions.

However, the techniques are not exclusive but are combined according to the type of application. The next stage is the description or extraction of features, it consists of extracting features with some quantitative information of interest or that are fundamental to differentiate one class of objects from another.

Then, the recognition stage is the process that assigns a label to an object based on the information provided by its descriptors. Interpretation involves assigning meaning to a set of recognized objects. Finally, the Knowledge Base stage, which will store the problem domain to guide the operation of each processing module, also controls the interaction between modules [5]. We will explore fundamental methods in image processing. Therefore, the following concepts will be defined:

Thresholding: Thresholding is one of the most important methods of image segmentation. The threshold is defined as a function that converts an image with different shades to a black and white image. If the original image is $f(x, y)$, the umbralized image is $g(x, y)$ and a threshold U is set $(0 < U < 255)$, the thresholding operation is defined as:

$$g(x,y) = \begin{cases} 255, & \text{if } f(x,y) \geq \text{Threshold.} \\ 0, & \text{if } f(x,y) \leq \text{Threshold.} \end{cases} \quad (1)$$

A threshold is selected that allows the pixels of an image that belong to various objects in the same image to be grouped, differentiating them from the background. In this way, histogram-based segmentation is based on the choice of one or more thresholds. These thresholds allow the points of the image to be grouped into regions of similar characteristics based on their gray levels [5].

Gabor Filter: Noise in an image refers to unwanted and unpredictable variations in the intensity levels of pixels. Applying a filter or mask to an image is to highlight or smooth out specific features. Gabor filters are used for texture analysis and edge detection in images. They are designed to capture features at different frequencies and orientations. The Gabor function is a combination of a sine wave modulated by a Gaussian function.

Otsu Method: Is a technique used in image processing to perform automatic thresholding. An image is divided into two distinct classes ("target" pixels and "background" pixels) based on a threshold value. The objective of the Otsu method is to find the optimal value of the threshold that maximizes the variance between the two resulting classes.

Segmentation: Subdivides an image into its constituent parts or objects, in order to separate the parts of interest from the rest of the image, therefore the level at which this subdivision is carried out depends on the problem to be solved. The basic segmentation attributes of an image are: luminance in monochromatic images, color components in color images, texture, shape, etc. Monochromatic image segmentation algorithms are generally based on one of the two basic properties of gray level values: discontinuity and similarity. In discontinuity, the method consists of dividing an image on the basis of sudden changes in the gray level.

The most important issues in discontinuity are: a) detection of isolated points, b) detection of lines and c) detection of edges of an image. In the similarity, the regularity in the gray level values is presented, the main methods are based on thresholding, region growth, and division and fusion of regions [2].

Edge Detection: Although the detection of dots and lines is essential in any presentation of image segmentation, the most common technique for identifying significant discontinuities in gray levels is edge detection. This is because these discontinuities are more frequent in practical applications. The methods of extracting edges from an image are based on the differences experienced by a feature between two adjacent regions, signaling the presence of an edge. In general, different models of edges or contours are identified: line, step type, ramp type and roof type. Discontinuities are detected using first- and second-order derivatives, in the case of first-order derivatives the gradient operator is used, while in second-order derivatives the Laplacian operator is used [5]. The first derivative produces a highlight of the areas in which the intensity is not homogeneous. The second derivative causes a sign change at the edge position.

These are the most common operators to detect edges are the following: canny, roberts, sobel and laplacian.

Edge binding and boundary detection: While there are several techniques that detect intensity discontinuities, and that should result in pixels that are at the boundary between an object and its background, in practice, this set of pixels rarely characterizes a boundary completely. This is due to noise, interruptions at the boundary due to non-uniform illumination, and other effects that introduce spurious intensity discontinuities. Therefore, edge detection algorithms are followed by a union and other border detection procedures designed to gather the edge pixels into a meaningful set of object borders [2]. Below are some techniques that fit that goal:

– *Local processing*: It consists of analyzing a neighborhood environment (for example, $3x3$, $5x5$) on all the points (x, y) of an image in which an edge detection process has been carried out, so that all the points that have similar characteristics join, forming a common border. Two properties are used to determine the similarity between edge pixels: a) Magnitude of the gradient vector (threshold value to determine the edge), and b) Direction of the gradient [5]. The first property states that a coordinate pixel (x, y) is similar to another (x, y) within its neighborhood environment if the equation holds:

$$[F(x, y)] - G[F(x', y')] \leq T, \tag{2}$$

ehere T is a non-negative threshold.The second property, i.e. the direction of the gradient (G), can be established using the angle of the gradient vector (θ) which is given by:

$$\theta = tan^{-1} \frac{G\mathbf{x}}{Gy}, \tag{3}$$

where (θ) represents the angle with respect to the x-axis. Thus, an edge pixel located at (x, y) has an angle similar to (x, y) which is:

$$|\theta - \theta'| \leq A. \tag{4}$$

where A is an angular threshold and (θ) is the angle of the gradient vector. A point in the predefined neighborhood of (x, y) is bound to the pixel of (x, y) if the magnitude and direction criteria are satisfied. It is repeated for each position of the image. A record should preserve the linked points as the center of the neighborhood moves from pixel to pixel.

– *Global processing using the Hough Transform*: Originally it was designed to detect lines and curves, using known analytical equations of object edges, however with this original method, it is not always possible to find analytical equations describing edges. The generalized Hough Transform makes this possible, i.e. to detect edges of objects even when the edge analytic expressions are not known. Intuitively, this method of edge detection consists of calculating the gradient of an image, then creating an accumulation field based on the parameters of the function being searched, and subsequently

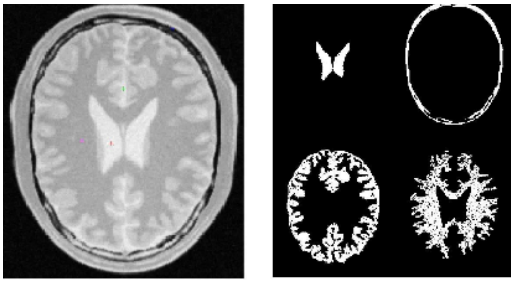


Fig. 2. Segmentation by region growth.

the maxima in the accumulation space indicate the existence of the searched objects. Previously, before applying the transform, the input image will be the binary image of the pixels that are part of the image contour. It allows to detect edges of images that are not lines, circles or ellipses. Likewise, it allows detection of objects with predetermined shapes, it is even possible to detect objects whose exact shape is not known but having or assuming prior knowledge an approximate model of the object can be formed [5].

Region-Oriented Segmentation: The spatial characteristics of an image are used to carry out its segmentation into regions. In other words, the image is divided into related regions, where each region has distinct properties that differentiate them from each other. In short, the goal is to extract the objects from an image, allowing their processing to be independently. Several techniques have been developed for this classification, and in this paper two of the most commonly used are presented: a) the growth of regions and b) division and fusion.

- *Growth of regions*: The region growth method is widely used and is based on the grouping of adjacent pixels that share similar characteristics or properties. This process starts with a set of points called "seed" and the regions are grown by incorporating neighboring points that exhibit similar properties, such as intensity, texture, color, among others. For example, if the considered property is intensity, a common criterion for including a pixel in a region might be that the absolute difference between the intensity of the pixel and the intensity of the seed is less than a predefined threshold [2, 7]. In Figure 2 the process is observed, on the left side of the image, the seed pixel labeling is found, while on the right side the region growth technique is shown using the increasing region brightness rule.

- *Division and merger of regions*: It consists of initially dividing an image into a set of arbitrary disjoint regions, for example 64 divisions, then, depending on the segmentation criterion, adjacent regions are merged if they have similar properties such as similar gray level, or are divided if they do not share the same properties, such as considerable variations of gray levels.
  Finally, the image is segmented into a set of homogeneous regions [6, 9].

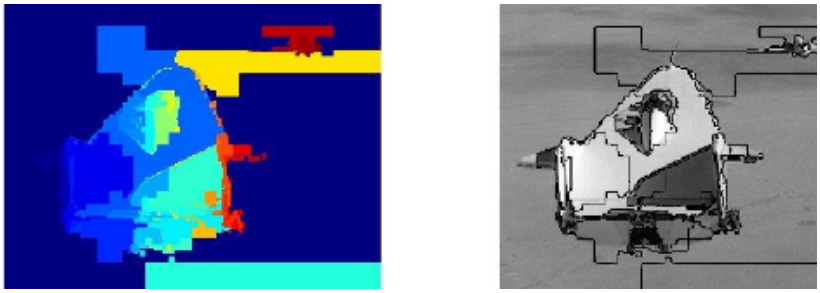


Fig. 3. Stages of image processing.

Figure 3 shows this process: on the left is the labeled image and on the right, the segmented image using this technique. One observation about this technique is that it does not preserve the actual contours in the result.

1.1 Related Works

In image processing, the continuous nature of data makes gradient-based attack approaches applicable, making it possible to create subtle and specific disturbances in images. However, the detection and classification of urban and rural areas present particular challenges due to the variability in visual and spatial characteristics. In this section, we present works related to the analysis and processing of images for the identification and characterization of urban and rural areas using different techniques and architectures:

In [8], image fusion techniques are presented to improve and calibrate meteorological information from ground-based radar images using satellite images. To calibrate the radar images, a method based on the discrete wavelet transform was implemented. A methodology is defined for the selection of global segmentation thresholds and for the calibration of radar images.

In [1], aims to show the results of the application of two cloud segmentation techniques in GOES satellite images; the first is a region-based technique, the thresholding by gray levels; and the second is a border-based technique, the Hough Transform. Finally, the results found by the two methods are compared with the segmentation obtained from a software specialized in satellite images by separating the spectral band with the information of interest.

In [3], the objective is to establish a reliable methodology of automatic detection of buildings for the automatic classification of land uses in urban environments using high resolution aerial images and LiDAR data. These data correspond to the information acquired with in the framework of the National Plan of Aerial Orthophotography (PNOA), and are available to the Spanish public administrations.

In this paper we analyze the development of an advanced satellite image segmentation system in urban and rural environments. Specialized filters are applied to eliminate noise and improve the quality of the images. The

Otsu method adjusts the threshold to improve the segmentation of regions, identifying homogeneous areas accurately. Gabor filters highlight important textural features, improving the distinction of details in complex environments. Segmentation techniques based on regions and edges are combined to obtain accurate and detailed results. The parameters of the Hough transform are optimized to reliably detect boundaries and contours, accurately representing the geospatial structure. This approach contributes to applications such as urban planning, agricultural monitoring and environmental change analysis, providing detailed information on urban and rural characteristics in complex spatial contexts.

2 Methodology

In the development of this work, satellite images covering urban areas, roads, uninhabited areas and vegetation in PNG format were used [4]. These images were subjected to segmentation methods, the first of them based on regions. Given the distinctive characteristic of vegetation, it was chosen to use the gray level thresholding method, which is equivalent to a segmentation by division and fusion of regions. On the other hand, the second selected method is based on the identification of limits, and for this the Hough transform was applied.

The edges of the images were extracted and subsequently compared with the results obtained by the aforementioned methods. To evaluate the effectiveness of the methods, metrics such as the mean square error, calculated pixel by pixel, as well as the Jaccard Index (IOU) and the Dice coefficient were used. In the initial phase, image selections were made, followed by grayscale conversion and subsequent binarization of the same. Given the observation that testing with different thresholds would take a long time, the decision was made to employ the Otsu method. This method made it possible to find the optimal threshold in an automated way in relation to the image in question, streamlining the segmentation process.

Subsequently, Gabor filters were used in order to enhance certain textures or specific characteristics. While these filters are effective at highlighting fine details, they also have the potential to introduce some level of noise into the image. To counteract this unwanted effect, noise filtering techniques were implemented, including the median filter, average filter, Gaussian filter and bilateral filter.

In this instance, the decision was made to maintain the result obtained with the bilateral filter. Although the changes were not obvious to the naked eye, a subtle alteration in the frequency histogram of the image could be perceived. This adjustment was crucial, since a slight noise had been generated in the image that, although it was not visually perceptible, was reflected in the histogram analysis.

Then, edge detection was carried out, using operators such as Canny, Laplacian, Roberts and Sobel. This procedure aimed to evaluate which of these operators could offer optimal performance. After edge detection was

completed, region-oriented segmentation was performed. In this case, we chose to use the Felzenszwalb method, an algorithm designed to divide an image into homogeneous regions. This method is based on the identification of natural boundaries in the image and the grouping of nearby pixels that share similar characteristics. The following steps describe the operation of the algorithm:

- *Construction of the Graph*: The image is represented as an undirected graph, where each pixel is a node and the connection between pixels is established according to some measure of similarity, such as the difference in color or intensity.
- *Ordering of the Edges*: The edges of the graph are ordered according to the similarity measure. This is done in an ascending manner, so that the edges with the least similarity are considered first
- *Hierarchical Grouping*: The edges are traversed in ascending order and the nodes connected by each edge are grouped if they belong to different regions. Grouping is done using efficient data structures, such as union-find.
- *Binding Criterion*: The decision whether or not to join two regions is based on a threshold criterion that compares the similarity measure between the pixels connected by an edge with a predefined threshold value.
- *Adjusting Segment Size*: A parameter is entered that controls the minimum size of a segment. This parameter influences the joining decision, allowing small regions to be merged or separated according to the desired size.
- *Linear Complexity*: An outstanding feature of Felzenszwalb's method is its linear complexity. As the edges are processed in ascending order, the time complexity is proportional to the size of the graph, which makes it efficient.

The final result is a partition of the image into homogeneous regions, where the natural boundaries have been respected to the extent that the similarity between connected pixels is greater than the defined threshold. Felzenszwalb's method is effective in segmenting images with complex structures and local variations in texture and color. The variation of these parameters can have the following effects on the scale adjustment: a) Increasing, generates larger segments and less details, which can lead to a more generalized segmentation, and b) Decreasing, generates smaller segments and more details, which can result in a finer and more detailed segmentation.The parameters that were modified from this method are as follows:

- *Standard Deviation Adjustment (sigma)*: Increasing sigma, the segmentation becomes less sensitive to intensity variations, resulting in a smoother segmentation. Decreasing sigma, the segmentation becomes more sensitive to intensity variations, resulting in a finer and more detailed segmentation.

– *Minimum Size Adjustment (min-size)*: A larger value, there is a larger merging of small segments, resulting in a smaller number of segments in the output, and a smaller value, there is a smaller merging of small segments, resulting in a larger number of segments in the output.

Then, to address edge-based segmentation, the Hough Transform was implemented with the specific objective of delimiting roads or tracks. We chose to use the implementation of the probabilistic Hough Transform in the Python CV2 libraries. This variant represents an optimization of the original Hough Transform. Instead of considering all the points in the image, the probabilistic Hough Transform selects a random subset of points, which is sufficient for line detection. This strategy contributes to an improvement in computational efficiency. It is only necessary to reduce the threshold to adapt to the use of this subset of points, thus simplifying the line detection process.

Finally, the implemented methods were evaluated by applying specific metrics. Key measures such as mean square error, Jaccard index (IOU) and DICE coefficient were used. These metrics provide a quantitative evaluation of the performance of the segmentation and detection methods, allowing an objective and detailed comparison of their results. The metrics were computed as follows:

– *Mean Square Error (MSE)*: measures the average quadratic difference between the actual and predicted values. The lower the MSE, the better the agreement between the predictions and the current values:

$$MSE = \frac{1}{N}\sum_{i=1}^{n}(y - \hat{y})^2. \tag{5}$$

– *Jaccard Index (IOU)*: measures the similarity between two sets. In the context of image segmentation, A and B represent the segmented and reference areas, respectively:

$$IOU = \frac{|A \cap B|}{|A \cup B|}. \tag{6}$$

Where $A \cap B$ it is the size of the intersection between the segmented image (A) and the reference image (B), and $A \cup B$ is the size of the junction between the segmented image (A) and the reference image (B).

– *Coefficient DICE*: Is a measure of the similarity between two sets. Similarly to the Jaccard Index, it is commonly used in image segmentation evaluation. The closer the DICE coefficient is to 1, the greater the agreement between the segmented and reference areas:

$$DICE = \frac{2 \cdot |A \cap B|}{|A + B|}. \tag{7}$$

Where $A \cap B$ it is the size of the intersection between the segmented image (A) and the reference image (B), and $A + B$ is the sum of the size of the segmented image (A) and that of the reference image (B).

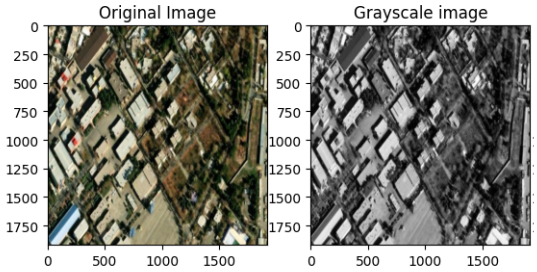


Fig. 4. Original image and with the pre-processing from color scale to grays.

The choice of these metrics is crucial to understanding the effectiveness of the methods in terms of accuracy, robustness and ability to capture the relevant information in the images. The mean square error provides a measure of the discrepancy between the segmentations and the reference images. On the other hand, the Jaccard Index and the DICE coefficient evaluate the overlap between the segmented areas and the reference areas, which are valuable indicators of the quality of the segmentation in terms of concordance with earthly truth.

3 Results

This section presents the results obtained from the various experiments carried out in this practice, along with their analysis and comparison. From the set of available images, 10 test images were selected for further analysis. The first stage of the process consisted of opening an image for manipulation, which has a size of 1920 x 1920 pixels, followed by its conversion to grayscale. A representative example of the results obtained is shown in Figure 4, corresponding to image number 4 indicated in Tables 1 and 2.

As detailed above, the next phase of the procedure involves the binarization of the image, i.e. the application of a threshold to distinguish the areas of interest. In this context, the Otsu method was implemented to optimize the assignment of this threshold automatically, using the OpenCV library. This approach seeks to improve the efficiency of the process by automatically determining the most appropriate value. Through experimentation, a significant distinction was made between manual binarization and the application of the Otsu method, which optimally determined that the most suitable threshold for the image in question is 111, as shown in Figure 5. Next, a Gabor filter was implemented with the purpose of highlighting textured areas or elements that present finer details. The hyperparameters used were: a kernel size of 31, a sigma value of 4.0, a theta value of 10 (filter orientation), a lambda value of 10 (filter wavelength), and a gamma value of 0.5 (filter aspect factor). This process is illustrated in Figure 6. Then a filtering was applied to eliminate the noise that could have been generated in the image by using the Gabor filter. Several filters were analyzed, including the median, average, Gaussian and bilateral. A 5×5 kernel was used for the Gaussian filter and a 3×3 one for the medium one.

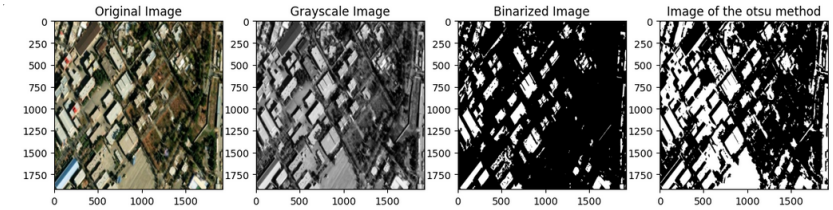


Fig. 5. Comparison with normal binarization and with the otsu method.

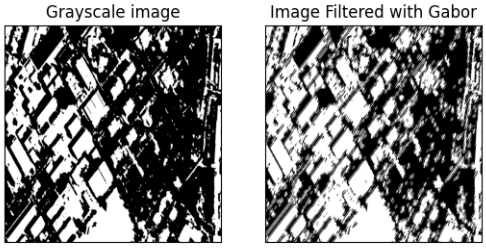


Fig. 6. Image with binarized pre-processing and with gabor filter.

At first glance, there were no significant changes in the application of the various filters. To select the filter that would be kept for further work, their histograms were plotted and the variations between them were evaluated. Although a subtle difference was observed between the filters, it was decided to work with the bilateral filter due to its ability to preserve the image characteristics while providing effective smoothing. After choosing the bilateral filter, we proceeded to extract the edges of the image by applying operators such as Canny, Sobel, Roberts and Laplacian. The results obtained with each of these operators are presented in Figure 7.

After completing this process, the analysis was initiated by region-oriented segmentation, specifically using the gray-level-based division and merging of regions approach. The algorithm mentioned above in the introduction was implemented, which is typical of the scikit-image library (skimage). This algorithm is fundamentally based on Canny's edge operator and was configured with specific parameters, including scale= 70, sigma= 0.5 and min-size=50. The result is shown in Figure 8.

This operation was performed on several images, and metrics such as the Jaccard Index (IOU) and the DICE coefficient were used to evaluate the effectiveness of the process and quantify the number of segments generated in each image. The detailed results are presented in Table 1. To analyze these metrics, ground truth (the reference image in gray scale) and image segmented by this method were used to analize the effectiveness of obtaining the correct segmentation.

As we observed in the results, images 31, 34, and 50 seem to have very good results, and the current thresholds could be considered effective. It is important to consider the specific context of your data and your goals to determine which

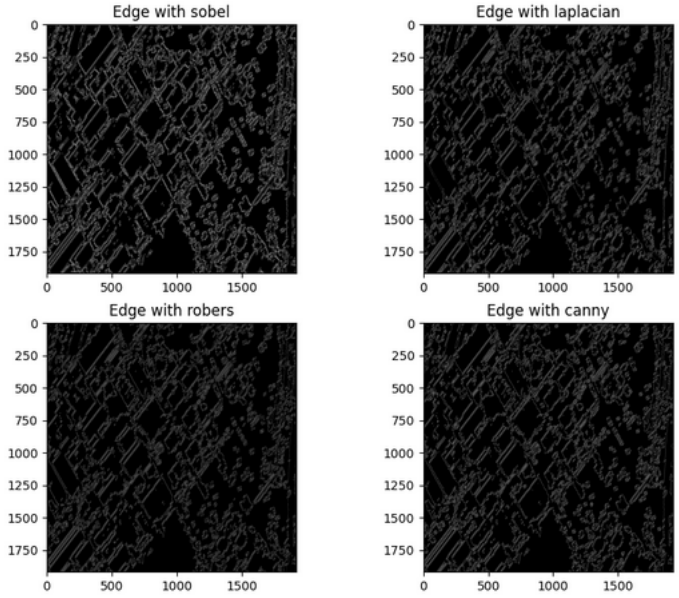


Fig. 7. Operators for edge detection.

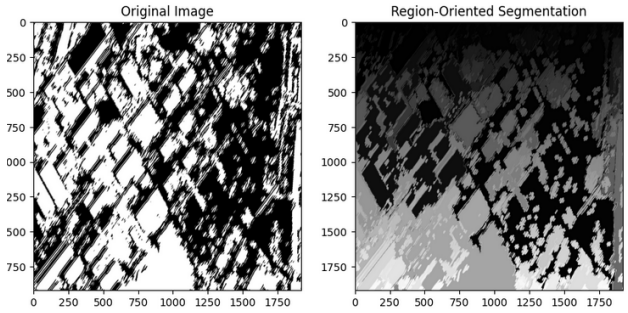


Fig. 8. Segmentation based on division fusion of regions by gray levels.

results are considered acceptable or desirable in your particular application. In addition, the interpretation of the metrics may vary depending on the specific requirements of the project.

Now we will proceed to analyze the border-based segmentation using the Hough transform that we described above. The initial parameters with which we trained was using the Canny edge operator, a threshold of 179, line of 50 and gap of 50, the results are shown in Figure 9. According to the results shown in Figure 9, it is evident that when using this operator, the image presents some distortion. The lines, which are thick, make the image choppy, mostly segmenting large areas.

Although it manages to segment some crossings, it does not do it optimally. Lowering the threshold only results in thinner lines without substantially

Table 1. Results of segmentation based on division and mergers of regions by gray levels.

ID-image	Threshold	IOU	Coefficient DICE	N-regions
4	111	0.9999	0.0018	3370
15	111	0.9957	0.0009	5354
28	114	0.9349	0.0015	3938
31	89	0.9978	0.0029	2792
34	90	0.9998	0.0030	1804
37	102	0.7417	0.0022	3581
42	103	0.8212	0.0026	2943
50	108	0.9859	0.0052	2055

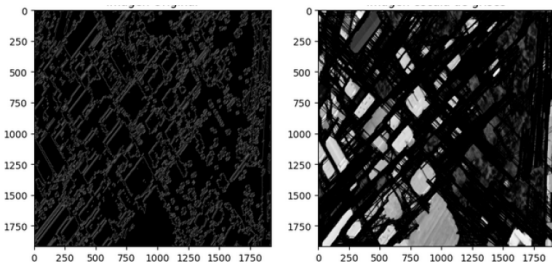


Fig. 9. Border-based segmentation using hough transform.

improving the quality of the segmentation. Therefore, the test was carried out with the Sobel, Roberts and Laplacian operators to analyze the differences. When analyzing the images, a remarkable similarity was observed between the three. The difference lies in the choice of the operator and the application of a threshold of 100, a line of 50 and a space of 50. This adjustment allowed a more effective segmentation of the edges or boundaries of the elements present, such as houses, roads and non-populated areas, while the rest corresponds to the vegetation zone, see Figure 10.

In Table 2 the mean square error of the 10 images was analyzed with each of the operators.To do this, the segmentation lines were drawn on the original grayscale image, and the error was calculated between that image and the image with the detected edges of all the elements to check the effectiveness of the segmentation.

According to the results of Table 2, the average and the standard deviation of each operator were analyzed. The results are as follows: *Sobel* (Average 77.7451, Standard Deviation 9.3545), *Canny* (Average 85.8375, Standard Deviation 23.2516), *Roberts* (Average 60.7196, Standard Deviation 18.2048), *Laplacian* (Mean 60.9174, Standard Deviation 16.3434). Observing these results, Sobel and Laplacian have lower standard deviations, indicating less variability in edge detection. Therefore, more robust options against noise could be considered.

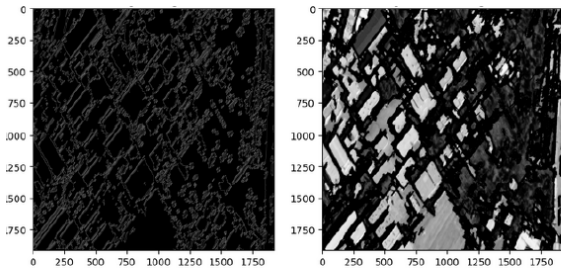


Fig. 10. Edge-based segmentation by Hough transformation using Sobel.

Table 2. Results of the mean square error of boundary-based segmentation using hough transform.

ID-image	Sobel	Canny	Roberts	Laplacian
4	73.9297	84.1937	53.6348	52.6819
15	63.707	39.0652	28.9142	35.4735
28	66.7526	63.3363	43.2689	43.2352
31	75.4467	93.9469	58.9237	58.3859
34	86.6452	106.3451	75.0496	74.0103
37	82.3606	99.3396	68.2301	67.2245
42	84.9869	98.7343	74.9895	74.3306
50	90.1322	101.7385	82.7425	81.9969

4 Conclusions

The choice between the two types of segmentation, based either on the division and merging of regions or focused on borders, significantly depends on the specific objectives of the analysis. In the case of merging regions based on gray levels, it is observed that it allows to detail areas with greater precision, which is beneficial when you want to highlight specific aspects of the image. On the other hand, border segmentation facilitates the identification of segments that act as borders, being useful, for example, to mark areas of interest in satellite images, such as populated areas, homes, roads and uninhabited regions.

It is noteworthy that, in the analysis of the Hough transform, it was evident that, unlike many works that use the Canny edge detector, in this case, methods such as Sobel, Roberts or Laplacian offered more satisfactory results by highlighting the relevant contours and features more effectively. In addition, Gabor filters were introduced to emphasize textures and differentiate features such as vegetation and empty areas. It is essential to recognize that the choice of each method and the combination of techniques will depend on the specific nature of the images and the particular objectives of the analysis. As a future work, applying deep learning methodologies would help the segmentation processes and adapt to more complex patterns in images. Consider the integration of multispectral data to take advantage of, in order to improve the accuracy of the segmentation of features such as vegetation and water. The

analysis of temporal changes in images could also be addressed, allowing the identification of variations in the landscape over time.

Acknowledgments. The first author thanks the support provided by the CONAHCYT scholarship number 1106756.

References

1. Gómez Vargas, E., Obregón Neira, N., Rocha Arango, D.F.: Cloud segmentation methods applied to satellite images. Tecnura 17(36), 96–110 (2013)
2. Gonzales, R.C., Woods, R.E.: Digital image processing. pp. 567–634. Prentice-Hall (2002)
3. Hermosilla Gomez, T.: Detección automática de edificios y clasificación de usos del suelo en entornos urbanos con imágenes de alta resolución y datos LiDAR. Ph.D. thesis, Universitat Politècnica de València (2011)
4. Moradi, S.: Segmented satellite images of buildings (2023), https://www.kaggle.com/datasets/sohaibmoradi/segmented- satellite-images-of-buildings
5. Palomino, N.L.S., Concha, U.N.R.: Segmentation techniques in digital image processing. Revista de investigación de Sistemas e Informática 6(2), 9–16 (2009)
6. Platero, C.: Computer Vision Notes Ch. 5, ISA, UMH (2007), https://isa.umh.es/asignaturas/rvc/cap5VASegmentacion.pdf
7. Reid, M., Millar, R.J., Black, N.D.: Second-generation image coding: An overview. ACM Computing Surveys (CSUR) 29(1), 3–29 (1997)
8. Socorras Quintero, V.S., Gómez Vargas, E., Obregón Neira, N.: Calibration of weather radar images. Tecnura 18(41), 12–26 (2014)
9. Sonka, M., Hlavac, V., Boyle, R.: Image processing, analysis and machine vision. pp. 112–192. Springer (1993)

ISSN 1870-4069

Multimodal Misinformation Detection from YouTube Videos Employing on Early and Late Fusion

Luz Elisa Gahona-Castillejos, Jesús Ariel Carrasco-Ochoa, José Francisco Martínez-Trinidad

Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico

{luz.gahona, ariel, fmartine}@inaoep.mx

Abstract. Misinformation has gained importance recently as online platforms and social networks have become so popular and influential in daily life, such as elections, public health, and the economy. Misinformation poses a challenge, and current approaches have not yet produced an effective solution. In this paper, we propose to focus on detecting multimodal misinformation in YouTube videos. Our proposal involves extracting features from text and audio independently and then combining them. Traditional features such as bag-of-words and embeddings were used to represent the text extracted from videos, and we also explored pre-trained transformations. Regarding the audio extracted from videos, a windowed feature extraction method was proposed, which includes zero crossings, root mean square (RMS), and eGeMAPS features. The best features of both modalities were combined in an early fusion, and then late fusion was used to improve misinformation detection accuracy, as shown in our experiments.

Keywords: Bag of words, transformes, RMS, models, fusion late, misinformation.

1 Introduction

The term "disinformation" refers to information that is both false and intentionally disseminated to mislead and harm others. This harmful intent is a crucial aspect of disinformation, yet it is often overlooked. In contrast, "misinformation" also involves spreading false information without the intent to cause harm [2].

In recent years, the spread of fake news and harmful language on various online platforms, especially social media, has become a significant concern. Detecting and curbing the dissemination of such content has become a priority for researchers and governments [3].

Misinformation detection is also related to fact-checking since both tasks aim to assess the veracity of claims [6]. Depending on where the information is

sourced. It may include text only, or audio only in the case of spoken interviews. Multimodal data is also common when sourced from videos, as it can consist of audio, and text when the video is transcribed.

The speech signal is complex as it carries information about the message content, speaker, language, and emotions. While many speech-processing systems work well with neutral speech in controlled environments, they often struggle with emotional speech due to the complexity of modeling and describing spontaneous speech.

The meaning of a spoken text can change based on intonation and context. For instance, "OKAY" in English can express emotions and attitudes like admiration, disbelief, consent, disinterest, or affirmation.

Therefore, understanding the text alone cannot fully capture a spoken statement's meaning [8]. The problem of misinformation has been addressed in different ways, one of them being during the COVID-19 pandemic, when a lot of misinformation was spread on YouTube and other social networks. [12] proposes a simple methodology based on NLP that can help fact-checkers detect misinformation about COVID-19 on YouTube, this work focuses on the comments of the videos (a single modality) from which they extract features that can be used for misinformation detection and creates a multi-label classifier based on transfer learning that can detect misinformative comments.

Accurate detection of misinformation content is essential to mitigate its impact and promote a more reliable and truthful information environment. By combining text and audio modalities, the characteristics of what is said and how it is being said could help in detecting misinformation more effectively.

Following this idea, in this paper, we propose a multimodal approach for detecting misinformation based on early and late fusion.

The paper is structured as follows. Section 2 reviews the related works closest to our proposal. Section 3 presents our proposal for detecting multimodal misinformation, detailing how to handle both modalities through early and late fusion. Section 4 shows the experiments carried out to validate our proposal. Finally, Section 5 includes our conclusions and future work.

2 Related Work

Previous research has explored the capabilities and limitations of NLP in identifying incorrect information. In [11], the authors propose focusing on generalization, and uncertainty and leveraging recent language models such as RoBERTa-large and GPT-4. On the other hand, [13] suggests a Multimodal Co-Attention Network (MCAN) that better fuses textual and visual features in fake news detection.

Another approach was proposed by [12], which uses YouTube video comments and pre-trained transfer learning models to generate a multi-label classifier that can categorize conspiratorial content. Today, one of the most used platforms is YouTube, where you can find a lot of misinformation about, health problems such as being overweight. Because YouTube cannot inform the user whether the

information in a video is true or not, there is a need to develop methods to recognize, from the videos, whether the information is correct.

Most misinformation detection work focuses on the text modality, with data extracted from platforms such as Twitter.

We propose from YouTube videos, taking only the audio and transcribing, this detection can be carried out.

Although the following works do not focus on the detecting of misinformation, they do focus on the recognition of emotions in a multimodal way using audio and its transcription, offering an idea of how to address misinformation detection. In [5] uses BERT, which is known to be very effective for many text classification tasks, and Mel coefficients for audio characterization. Different approaches have been proposed. For example, in the multimodal system developed by Robinet [7], the Inception-ResNet-V2 neural network is used for emotion recognition in speech. This network receives the audio's melodic frequency cepstral coefficients (MFCC) as input and RoBERTa is used to extract features from the text modality. The fusion of modalities is carried out through a closed attention mechanism.

On the other hand, in [9], the extraction of text features is addressed using the GloVe and Elmo embedding models. For audio, a convolutional neural network (CNN) is used along with a long-term recurrent memory (LSTM) model.

The features extracted from both modalities are concatenated for further processing. [7] and [9] works utilize the melodic frequency cepstral coefficients (MFCC) as the primary audio characteristic, obtained by segmenting the audio into windows. These coefficients serve as input for the respective neural networks. Another appears feature used in voice-related work is the root mean square (RMS).

Different types of fusion combine two or more modalities depending on the stage they are combined. Early fusion, also known as data-level fusion or feature-level fusion, occurs at a very early stage of model development, before any network layer. It appears at the raw or preprocessed data stage, or simply when features have been extracted from the raw data. Late fusion, also known as decision-level fusion, involves first developing a complete model for each modality. The individual models' outcomes (decisions or probabilities) are then integrated [3].

3 Proposed Approach

As can be seen, there are currently no existing methods for detecting multimodal misinformation. Therefore, we propose a new approach for multimodal misinformation detection in this paper, based on early and late fusion.

As seen in Figure 1, the transcriptions are extracted for the text modality, and the audio from the video must also be extracted to process the voice signal; both modalities are used to detect misinformation from YouTube videos. In the following subsections, we describe each stage shown in Figure 1.

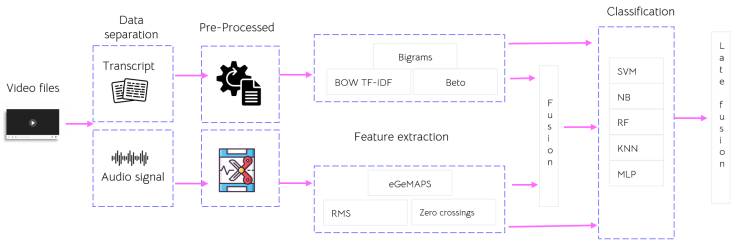


Fig. 1. Diagram of our proposal for Multimodal Misinformation Detection Employing Early and Late Fusion.

3.1 Data Separation

During this stage, we automatically obtained video transcripts to make it easier to analyze the spoken content. We also extracted the audio from the original video so that we could focus on the voice signal. This extraction enabled us to conduct a more detailed analysis of both the text and the audio.

3.2 Pre-processing

For the transcriptions, a cleaning process was carried out that included eliminating periods, commas, stop words, special characters, question marks, and exclamation marks; accents were left. Additionally, all text was converted to lowercase to ensure uniformity in the analysis. For the audio, some of them contained music, so a cleaning process was carried out to remove the music and leave only the voice. Subsequently, the audios were segmented into one-minute intervals. In cases where the audios were less than one minute, they were repeated until that duration was reached. The first minute of all the audios was taken for one set of experiments and, for another set, the minute from the moment speech was detected was used; this was done to the audios with and without music. Hamming windowing is applied, which avoids discontinuities introduced by analyzing only a fraction of the signal. The Hamming window function also has the property that its value varies smoothly from zero at the ends to one in the center, which helps avoid loss of information due to overlap between windows.

The Hamming windowing was applied to the audio with a size of $50ms$ per window and overlaps of $10ms$. The audios were also reduced to only the first 60 seconds; the rest of the audio was discarded from the analysis. If an audio did not reach that length, the signal was repeated until all had the same length, in normalize the sample size.

3.3 Feature Extraction

In this stage, the features of the video transcripts were obtained, which consisted of:

1. Bag of Words: This approach represent a document as a vector where each dimension corresponds to a single word in a predefined vocabulary. For weighing the words, we use the TF-IDF (Term Frequency-Inverse Document Frequency) by considering not only the frequency of words in the document, and their relative importance in the entire corpus.
2. N-grams: are contiguous sequences of n elements, which can be characters, words, or tokens. N-gram extraction allows for capturing contextual information by considering not only individual words, but also combinations of words that appear together. In our proposal, we use word bigrams (N=2).
3. BETO (Bidirectional Encoder Representations from Transformers for Spanish): BETO is an adaptation of BERT designed explicitly for the Spanish language. Like BERT, it uses a bidirectional transformer architecture to pre-train a language model on a large corpus of Spanish text.

The extraction of features in the audio part is for divided into two approaches. In the first, together with preprocessing, 13 Mel Frequency Cepstral Coefficients (MFCC) were extracted, where cepstral coefficients are a specific type of coefficients derived from cepstral analysis applied to a time window of the speech signal. This technique efficiently separates the two main components of information in the speech signal: excitation and vocal tract [10]. By zero crossings were obtained for each window, indicating how many times the speech signal crosses the zero level during a given period. This measure provides an overview of how the signal's energy is distributed [4]. Finally, the root mean square energy was extracted, where the energy of a signal refers to the sum of all the magnitudes of the signal. In the context of audio signals, this measure indicates the intensity signal's in terms of its volume or sound level.

The other approach consisted of using the Opensmile eGeMAPS feature set that consists of 88 static acoustic features resulting from calculating of various features on low-level descriptor functions, such as volume frequency loudness, alpha ratio, harmonic difference, mffc 1-4, loudness peak rate, among others [1].

3.4 Early Fusion

Once the individual features from each modality have been extracted, they are concatenated. For this experiment, we combined the features obtained from the text's term Frequency-Inverse Document Frequency (Tf-IDF) representation of the text with the individual features extracted from the audio signals. By integrating these features, we aim to leverage the strengths of both modalities: the semantic richness captured by Tf-IDF from the textual data and the emotional cues embedded in the audio features. This multimodal approach enhances the overall representation of the data, providing a more comprehensive basis for misinformation detection. The concatenation process involves aligning the feature vectors and creating a unified input for the classification model, thereby enabling the model to learn from text and audio characteristics simultaneously.

3.5 Classification

We trained several classifiers to detect misinformation the classifiers used were:
Linear SVM (Support Vector Machine): This classifier aims to find a linear hyperplane that separates the data into different classes in a multidimensional space. It is particularly beneficial when the data is linearly separable. By maximizing the margin between the classes, SVM ensures robust classification even with high-dimensional data.
Naive Bayes (NB): This classifer is based on the Bayes' theorem. It calculates the probability that a given instance belongs to a particular class based on the likelihood of observing certain features. Naive Bayes is known for its simplicity, efficiency, and effectiveness, especially with large datasets, and when the assumption of feature independence (naivety) holds approximately true.
Multilayer Perceptron (MLP): This is a basic form of artificial neural networks. MLPs are flexible and can learn complex relationships in the data through hidden layers. They are particularly effective in capturing non-linear patterns and interactions among features.
Random Forest (RF): This ensemble learning method is known for its performance on large datasets with many features. It is robust to missing data and outliers and does not require extensive data preparation. Random Forest builds multiple decision trees and merges them to get a more accurate and stable prediction.
K-Nearest Neighbors (KNN): This is a simple and effective instance-based learning classifier. KNN does not make any assumptions about the underlying data distribution and can capture complex relationships in the data. It classifies instances based on the majority label among the nearest neighbors, for our experiments we used K= 3.

3.6 Late Fusion

In addition to feature-level fusion, we implemented a late fusion strategy using five different classifiers: Support Vector Machine (SVM), Naive Bayes (NB), K-Nearest Neighbors (KNN), Multilayer Perceptron (MLP); and Random Forest (RF). Each classifier independently predicts the emotion class, and the final classification is determined by majority voting among the classifiers. This late fusion approach aims to improve the robustness and accuracy of the classification by combining the strengths of various classification algorithms.

4 Experiments

In the experiments, the data set was initially divided into 70% for training the models and 30% for testing. The IDs of both the audio and text elements were maintained in the same order to allow for proper analysis.

4.1 Dataset

Misinformation detection in the domain of Obesity and Overweight in Spanish language (MOOSP) is a dataset on obesity and overweight misinformation in Spanish, where queries such as "treatment for obesity and overweight", "risk factors for obesity", and prevention of obesity and overweight" were used to search for videos of interest. The database has a total of 243 videos, of which 113 are misinformation and 130 are informative, therefore the data collection is not imbalanced. The average length of the videos is 5.08 minutes, the most extended video in the database is 12.41 minutes, and the shortest one is 40.50 seconds.

The Language Technologies Research Group (LABTL) of Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE) developed the dataset used in this research. We thank Doctors Manuel Montes y Gómez, Luis Villaseñor Pineda, and Master Jennifer Pérez Santiago for providing access to these data, which were fundamental for the advancement of this study.

4.2 Text Experiment

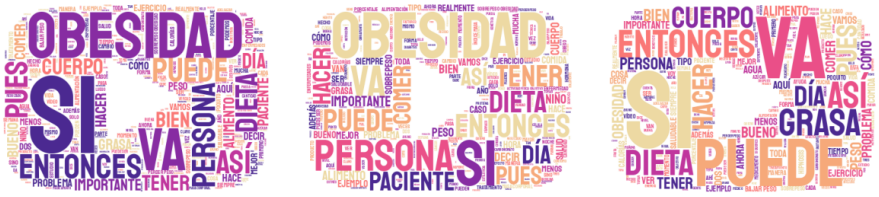
After obtaining the feature vectors, five different classifiers were used for the text modality. Each classifier processed the vectors individually to analyze the textual information. The majority voting combined the results of the five classifiers to ensure the precision and robustness of the predictions; this approach allowed for a final decision based on most individual predictions.

4.3 Audio Experiments

In the case of audio, several experiments were carried out to evaluate different scenarios. First, we worked with audio that contained music. From these audios, features from the first minute were extracted, and vectors were generated and then processed using the five classifiers mentioned above. This experiment was repeated by taking the audio sample from when speech was detected instead of the first minute. Subsequently, the same experiments were carried out with audio from which the music had been removed, leaving only the voice. In all cases, after the five classifiers analyzed the feature vectors, a majority voting was again applied to combine the results. In this way, it was possible to evaluate the influence on the performance and compare the results obtained with and without the presence of music in the audio.

4.4 Text Results

Once the corpus was pre-processed, the word cloud was obtained, which allows for identifying the most frequent words in a set of text quickly and easily. Each word is represented in the cloud with a size proportional to its frequency of appearance in the text, which can be helpful to obtain a quick understanding of the main themes or keywords present in the text; in the word cloud of the whole corpus you can see the filler words or statements such as "si", "va", This



(a) Word cloud of the corpus (b) Word cloud of the informative class (c) Word cloud of the misinformation class

Fig. 2. Word cloud of the corpus and the classes of informative and misinformation.

Table 1. Classification results (in terms of accuracy) with the different classifiers on the TF-IDF representation.

Classifer	Accuracy
SVM	0.55
MLP	0.55
Naive Bayes	0.58
Random Forest	0.49
KNN	0.52
Majority Voting	0.52

Table 2. Classification results (in terms of accuracy) with the different classifiers on the bigrams representation.

Classifer	Accuracy
SVM	0.79
MLP	0.78
Naive Bayes	0.78
Random Forest	0.74
KNN	0.71
Majority Voting	0.86

is seen in the Figure 2(a) while in the informative class words that stand out are "obesidad", "persona", and "dieta", see figure 2(b) as well as a more formal language, while in the word cloud of the misinformation class, more fillers are observed, the word "grasa" 2(c).

When reviewing the results in Tables 1-3 with the different types of features used, we can see that the representation of texts through Bigrams got the best accuracies. Neither TF-IDF nor Beto representations outperformed one of the results in Table 2.

Table 3. Classification results (in terms of accuracy) with the different classifiers on the BETO´s vector representation.

Classifer	Accuracy
SVM	0.27
MLP	0.52
Naive Bayes	0.68
Random Forest	0.40
KNN	0.51
Majority Voting	0.43

Table 4. Classification accuracy using eGeMaps for different classifiers without music.

Classifer	Accuracy
SVM	0.27
MLP	0.52
Naive Bayes	0.63
Random Forest	0.38
KNN	0.40
Majority Voting	0.39

4.5 Audio Results

After applying the methodology to the audio, the results without music are shown in Tables 4 and 5. In these tables, it was observed that the eGeMaps feature set was the most favorable. This finding suggests that the features extracted from eGeMaps provided relevant and discriminative information for classifying the audio into the desired classes. This result supports the effectiveness of eGeMaps as a set of features for detecting misinformation in audio.

For the audio data, as shown in Tables 4 and 5, the results obtained with the individual features were not very favorable for this modality. These tables included the audio without music, starting the analysis when the dialogue begins, as this approach provided better results compared to other representations, such as preserving the music or starting from the beginning of the videos.

4.6 Fusion

Using the results of the classifiers of the modalities that had the best results and then applying a majority vote, we obtain what is shown in the tables 6, which reflects an improvement compared to the unimodal result and their respective majority voting.

From the above experiments, we can appreciate that early feature concatenation effectively improved data representation. Subsequently, using majority voting in late fusion allowed for the outperformance of the individual modalities, reflected in a significant improvement in accuracy.

Table 5. Classification accuracy for different classifiers using RMS features without music.

Classifer	Accuracy
SVM	0.27
MLP	0.32
Naive Bayes	0.64
Random Forest	0.30
KNN	0.42
Majority Voting	0.34

Table 6. Classification accuracy of majority vote using eGeMaps and RMS with Bigrams.

Features	Classifer	Accuracy
EgeMaps + Bigrams	Majority Vote	0.85
RMS + Bigrams	Majority Vote	0.78

5 Conclusions

Detecting misinformation poses significant challenges depending on the analyzed data modality being analyzed. While textual features such as bigrams effectively capture nuances in text-based messages, acoustic features from eGeMaps emerge as a good option for misinformation analysis in audio formats. Integrating these modalities through concatenation does not necessarily improve misinformation detection accuracy, highlighting the complexities and potential redundancies introduced.

The results revealed that early feature fusion was particularly effective, providing a robust and significantly improved model performance. Subsequently, by applying late fusion, the individual modalities' performance surpassed the individual modalities' misinformation detection. From our experiments, We can conclude that our proposal captured contextual information, but also improved classification accuracy compared to models using only one modality. These findings underscore the importance of integrating multiple modalities and applying early and late fusion techniques to improve misinformation detection.

For future work, it is promising to explore using pre-trained language models, such as BERT or RoBERTa, for text feature extraction. These models have proven highly effective in various natural language processing tasks because they capture complex contexts and semantic relationships within text. Integrating these advanced features could significantly improve misinformation detection systems.

On the other hand, the application of transformers in audio characterization also deserves attention. Wav2Vec and HuBERT have shown great potential in extracting relevant features and improving tasks such as speech recognition, emotion detection, and acoustic event classification. By taking advantage of

transformers' capabilities, a richer and more accurate representation of audio signals can be obtained, which could lead to better results in various applications.

We are also considering exploring the possibility of adding the sequence of images from the video as another modality to analyze body language.

Acknowledgments. This research was partially supported by the National Council of Humanities, Sciences, Technologies, and Innovation of Mexico (CONA-HCyT) through its graduate study scholarship program.

References

1. Opensmile python documentation, https://audeering.github.io/opensmile-python/
2. Alam, F., Cresci, S., Chakraborty, T., Silvestri, F., Dimitrov, D., Martino, G.D.S., Shaar, S., Firooz, H., Nakov, P.: A survey on multimodal disinformation detection. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 6625–6643. International Committee on Computational Linguistics (2022)
3. Ayetiran, E.F., Özgöbek, O.: A review of deep learning techniques for multimodal fake news and harmful languages detection. IEEE Access 12, 76133–76153 (2024)
4. Bleda, S., Francés, M., Marini, A., Martínez, C.: Herramientas software para la docencia de la señal de voz en ingeniería técnica de telecomunicaciones (2024), https://ice.ua.es/es/jornadas-redes-2012/documentos/posters/246141.pdf
5. Das, M., Raj, R., Saha, P., Mathew, B., Gupta, M., Mukherjee, A.: Hatemm: A multi-modal dataset for hate video classification. In: Proceedings of the International AAAI Conference on Web and Social Media. vol. 17, pp. 1014–1023 (2023)
6. Hossain, T., IV, R.L.L., Ugarte, A., Matsubara, Y., Young, S., Singh, S.: Covidlies: Detecting covid-19 misinformation on social media. In: Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020. Association for Computational Linguistics (2020)
7. Khurana, Y., Gupta, S., Sathyaraj, R., Raja, S.: Robinnet: A multimodal speech emotion recognition system with speaker recognition for social interactions. IEEE Transactions on Computational Social Systems (2022)
8. Koolagudi, S., Rao, K.: Emotion recognition from speech: a review. International Journal of Speech Technology 15(2), 99–117 (2012)
9. Koromilas, P., Giannakopoulos, T.: Deep multimodal emotion recognition on human speech: A review. Applied Sciences 11(17), 7962 (2021)
10. Laynez, D.B.: Sistemas de Verificación Automática de Locutor, Capítulo 3. Proyecto fin de carrera, ingeniería superior de telecomunicaciones, Universidad de Sevilla, Sevilla, España (2012)
11. Pelrine, K., Imouza, A., Thibault, C., Reksoprodjo, M., Gupta, C., Christoph, J., Godbout, J.F., Rabbany, R.: Towards reliable misinformation mitigation: Generalization, uncertainty, and gpt-4. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 6399–6429 (2023)
12. Serrano, J.C.M., Papakyriakopoulos, O., Hegelich, S.: Nlp-based feature extraction for the detection of covid-19 misinformation videos on youtube. In: Proceedings of the 1st Workshop on NLP for COVID-19. Association for Computational Linguistics (2020)

13. Wu, Y., Zhan, P., Zhang, Y., Wang, L., Xu, Z.: Multimodal fusion with co-attention networks for fake news detection. In: Findings of the association for computational linguistics: ACL-IJCNLP 2021. pp. 2560–2569 (2021)

ISSN 1870-4069

A Proposal for the Recognition of Gait Pathologies in Individuals based on Multimodal Features

Iván J. Sánchez-Cuapio[1,3], Ricardo Ramos-Aguilar[2], Paola A. Niño-Suarez[1], Esaú E. Escobar-Juárez[2]

[1] Instituto Politécnico Nacional,
Escuela Superior de Ingeniería Mecánica y Eléctrica Unidad Azcapotzalco,
Mexico

[2] Instituto Politécnico Nacional,
Unidad Profesional Interdisciplinaria en Ingeniería Campus Tlaxcala,
Mexico

[3] Universidad Tecnológica de Tlaxcala,
Mexico

ivanjs24@gmail.com, pninos@ipn.mx, {rramosa,eescobarj}@ipn.mx

Abstract. This work focuses on developing a gait pathology diagnosis system using machine learning with multimodal features. The methodology combines inertial sensors and RGB-D cameras to understand gait patterns and identify movement pathologies in humans. The process begins with creating a controlled environment for data collection, using inertial sensors placed at key body points, such as the ankles, knees, and hips, along with strategically positioned RGB-D cameras. Data acquisition involves recording accelerations, rotations, images, and depth data during participants' gait. Subsequently, this data is preprocessed through cleaning, normalization, and noise removal, ensuring high-quality information. Key gait features, such as phase durations, range of motion, and oscillation frequencies, are extracted and used to apply multimodal classification algorithms like Random Forest, SVM, and CNN. These algorithms classify and diagnose gait pathologies based on temporal, spatial, and frequency characteristics. The methodology will be supported by publicly accessible databases such as the Daphnet Freezing of Gait Dataset, GaitRec Dataset, and Murooka Gait Dataset, which provide diverse data to validate and improve diagnostic models.

Keywords: Gait analysis, machine learning, multimodal features.

1 Introduction

Human gait is a cyclical process that allows human locomotion, characterized by a series of repetitive and coordinated movements of the lower limbs. Normal

gait involves a precise sequence of biomechanical and neuromuscular events that ensure stability, balance, and energy efficiency during locomotion.[19]. According to Perry and Burnfield [26], gait can be defined as a dynamic and complex process that involves the coordinated interaction of multiple body systems and is divided into two main phases: the stance phase and the swing phase, critical events that ensure stability, balance, and energy efficiency in the movement of people.

Gait problems in individuals are conditions that decrease the ability to walk normally and smoothly [27]. Gait pathologies, which refer to deviations from the normal pattern of locomotion, result from various neurological, musculoskeletal, and other disorders. According to Perry[26], these pathologies can be classified into several types, such as antalgic gait, Trendelenburg gait, hemiplegic gait, spastic gait, ataxic gait, neuropathic gait, parkinsonian gait, and foot drop gait. Kirtley emphasizes the importance of detailed and quantitative analysis to diagnose and treat these conditions, highlighting how gait pathologies can result from neuromuscular disorders, musculoskeletal injuries, and balance and coordination problems [19].

The development of technology to detect gait disorders offers significant advantages for patients, such as diagnosing with initial symptoms, continuous monitoring, objective evaluation, personalization, and assessment of treatment or therapy[6]. various technologies are used for gait analysis in individuals, some of the most common being force platforms, inertial sensors, electromyography, videography and image analysis, motion capture systems, and pressure insoles. The information from these systems is used in clinical research, evaluation, and analysis settings, allowing for the identification of abnormal patterns, performance assessment, and the design of rehabilitation interventions [9, 3, 11].

In gait analysis research, several studies have used innovative techniques with sensors and machine learning methods. Ionescu and Moga [17] present a gait recognition approach based on multiple projections and machine learning algorithms, highlighting the improvement in accuracy by combining different projections. Panwar and Gupta [25] review various gait recognition techniques using the Kinect sensor, discussing their effectiveness and challenges in capturing and analyzing gait data. Wang, Tan, Ning, and Hu [30] propose a gait recognition method based on silhouette analysis, applying machine learning algorithms for human identification. Chen, Jafari, and Kehtarnavaz [10] explore the fusion of depth and inertial sensor data for human action recognition, including gait, highlighting the improvement in recognition accuracy. Eskofier [13] discusses recent advances in the use of deep learning for sensor-based mobility analysis, emphasizing the integration of multimodal data for fall risk assessment. Zhang and Tao [33] introduce slow feature analysis for human action recognition, applicable to detailed gait analysis and capturing movement dynamics. These studies demonstrate the potential of multimodal technologies and machine learning to transform the analysis and evaluation of human gait.

The proposal in this work aims to achieve significant advantages over traditional approaches by integrating multiple data modalities, such as video images and inertial sensor data. This integration improves analysis accuracy

Table 1. Studies on Human Gait.

Autor-Published	Technology
Buffanti,2020	Camera RGB-D
Bijalwan,2021	IMU, Camera RGB-D
Palermo,2022	IMU, Camera RGB-D
Yamamoto,2022	IMU, Camera RGB-D
Cai,2023	IMU, Binocular Camera
Alanazi,2022	Camera RGB-D, Micro-Dopler
D'Antonio,2021	IMU, 3-WebCam
Albert,2020	Camera RGB-D
vanKersbergen,2021	Camera RGB-D

by providing a more comprehensive and detailed representation of human gait, mitigating individual errors from each sensor or modality. The combination of features captured by different sensors enables the detection of patterns to adapt to various conditions and environments. Altogether, the multimodal approach will capture subtle movement patterns and offer model adaptability in clinical, sports, and rehabilitation applications.

2 Background

2.1 Human Gait Parameters

Gait parameters are defined as quantitative measures used to describe and analyze human movement during locomotion. These parameters include kinematic, kinetic, and temporal variables that provide detailed information on how a person moves. Kinematic data describe the position and movement of joints and body segments in three-dimensional space during gait. This includes joint angles, range of motion, and movement patterns of each joint. Kinetic parameters quantify the forces and moments applied through the joints during ground contact, evaluating ground reaction forces, load distribution, and joint moments.

Finally, temporal parameters describe the duration of specific gait phases, such as stance time and swing time, providing information on the sequence and coordination of movement. These parameters are fundamental for understanding both normal gait and pathological alterations, allowing for a detailed analysis that guides the diagnosis and treatment of clinical conditions related to gait[19, 26]. Table 1 shows a summary of works related to gait analysis and the technology used.

Buffanti et al. demonstrate that non-invasive and cost-effective systems based on depth cameras can recover relevant features of human gait patterns. Gait data recordings were taken using multiple depth sensors. Time-domain analysis includes joint excursions across gait phases, range of motion (ROM), measures

of central tendency and dispersion, spatial variables, and center of mass (COM) position. Spectral analysis examines dominant frequency, magnitude, and phase shift during gait. Only features showing significant gender differences were used to train a Support Vector Machine (SVM) classifier [16].

Bijalwan et al. work on the biomechanics of pelvic, hip, knee, and ankle joint movements using a Kinect sensor and an inertial measurement unit (IMU) during normal walking. They present a cost-effective gait analysis system based on Microsoft Kinect v2 and an IMU device. The Kinect sensor is used to acquire 3D skeleton data (camera (x, y, z), depth (x, y), orientation (x, y, z, w), color (x, y)) with 25 human body joints. For their analysis, they consider lower limb joints, namely the spine joint, hip, knee, and ankle of both left and right legs [5].

Palermo et al. collect a multi-camera and multimodal dataset from patients walking with a robotic walker equipped with wheels and a pair of cameras. Depth data were acquired at 30 fps and synchronized with inertial data from Xsens MTw Awinda sensors and kinematic data from Xsens biomechanical model segments, acquired at 60 Hz [24].

Yamamoto et al. demonstrate the capability of markerless gait analysis using posture estimation based on a single RGB camera via OpenPose (OP) and an inertial measurement unit (IMU) on the foot segment to measure ankle joint kinematics under various walking conditions. Their proposed method has the potential to measure spatiotemporal gait parameters and lower limb joint angles, including ankle angles, as an assessment tool for gait in clinical environments [31].

Cai et al. present a procedure for joint angle estimation assisted by binocular camera to acquire initial orientations of the lower limb segment using a human pose estimation algorithm based on images and then estimate joint angle with kinematic constraint. The alignment procedure requires only a sitting posture and does not need any functional movement. Ten healthy participants were recruited for validation experiments, including standing up, turning around, and walking. The accuracy and efficiency of their alignment procedure were validated against optical motion capture (OMC) [7].

Alanazi et al. propose the use of millimeter-wave (MMW) radar as a promising solution for gait applications due to its low cost, improved privacy, and resilience to ambient light and weather conditions. They present a novel method of human gait analysis that combines micro-Doppler spectrogram and skeletal posture estimation using MMW radar, complemented by 3D coordinates extracted from 25 joints via Kinect V2 sensor [1].

D'Antonio et al. characterize the performance of a low-cost markerless system, consisting of the open-source OpenPose library, two web cameras, and a linear triangulation algorithm. The system was validated in terms of 3D gait kinematic analysis, compared with inertial sensors. They recorded synchronized videos of six healthy subjects in three webcam configurations, in walking and running sessions on a treadmill. They also compared sagittal joint angles between the two systems to assess the kinematic performance of the markerless system [12].

Albert et al. evaluate the motion tracking performance of the latest generation Microsoft Kinect camera, Azure Kinect, compared to its predecessor Kinect v2 in treadmill walking using a reference multicamera motion capture system Vicon and the Plug-in Gait model with 39 markers. Five young and healthy subjects walked on a treadmill at three different speeds while data were simultaneously recorded with all three camera systems. They used an easy-to-manage camera calibration method developed here to spatially align 3D skeleton data from both Kinect cameras and the Vicon system [2].

Van Kersbergen et al. studied the use of a depth camera to capture changes in the gait characteristics of Parkinson's patients. The dataset consisted of 19 patients (tested in both defined OFF and ON phases) and 8 controls, performing the "Timed-Up-and-Go" test multiple times while being recorded with the Microsoft Kinect V2 sensor. Derived features from the camera were step length, average walking speed, and mediolateral sway. Motor signs were clinically assessed using the Unified Parkinson's Disease Rating Scale by the Movement Disorder Society [18].

2.2 IMU Systems

Inertial systems are advanced technologies used to accurately capture and analyze parameters of gait and other human movements. These systems rely on sensors that measure linear acceleration and angular velocity of body segments. A typical architecture of an inertial system includes multiple sensors strategically distributed on the body, connected to a central processing unit that records and processes the data. Inertial sensors are small and lightweight, allowing comfortable and unrestricted data capture during gait. These systems provide precise measurements of kinematic parameters such as joint angles and movement trajectories, as well as temporal parameters like cadence, step length, and stance and swing times. This capability makes inertial sensors versatile tools in clinical settings for evaluating musculoskeletal disorders and in sports applications for performance analysis and functional biomechanics [28].

The optimal placement of inertial sensors for gait parameter recording depends on the biomechanical factors of human gait and joint movement, which affect the accuracy and reliability of collected data. Generally, it is recommended to mount sensors on body segments that undergo significant movements during gait, such as thighs, shins, and feet (Fig. 1). For example, placing sensors on the lumbar region or legs allows for direct capture of relevant joint angles and movement patterns. Moreover, precise placement at specific anatomical points, such as the anterior superior iliac spine for the pelvis or the knee center for knee joint flexion, ensures more accurate measurements of kinematic parameters. This strategy not only facilitates detailed assessment of gait biomechanics but also minimizes the risk of external interferences and motion artifacts, thus ensuring data quality for clinical analysis and sports applications [8].

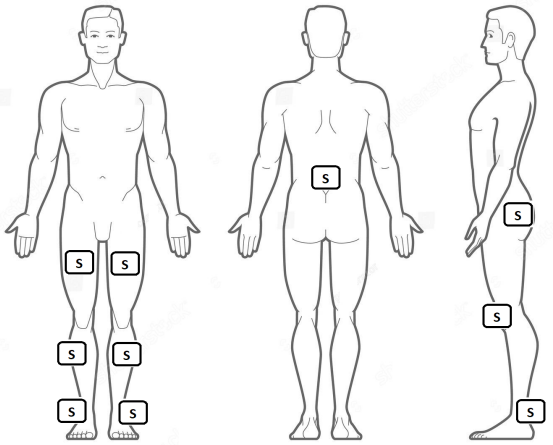


Fig. 1. Proposed placement of inertial sensors.

2.3 Camera with Depth

The use of RGB-D depth cameras, such as the Microsoft Kinect system, has revolutionized gait analysis by providing three-dimensional data capture that combines RGB sensors with depth sensors. This technology is known for its ease of use, minimal invasiveness, and cost-effectiveness, making it accessible in both clinical and research settings. Depth cameras allow for precise evaluation of kinematic and kinetic parameters without needing body-worn markers, thereby enhancing subject comfort. Their application in biomechanical research and rehabilitation has been extensively documented, highlighting their advantages and limitations compared to traditional motion capture systems. While they present challenges such as limited accuracy and dependence on lighting conditions, depth cameras offer a valuable tool for detailed and accessible analysis of human gait [29].

Proper placement of the capture system is crucial for obtaining accurate and reliable gait analysis data. The camera's location and angle determine the quality and precision of the measured kinematic and kinetic parameters. To achieve optimal motion capture, the camera should be positioned at an appropriate height and distance from the subject, typically at waist height and approximately 2-3 meters away (Fig.2). This positioning ensures that the subject's entire body is within the camera's field of view throughout the complete gait cycle. Additionally, adjusting the camera's tilt angle is important to maximize the visibility of body segments and minimize marker occlusion [14].

3 Datasets

In the context of gait analysis, the use of established databases such as GaitRec, MotionSense, and the CMU Graphics Lab Database is fundamental

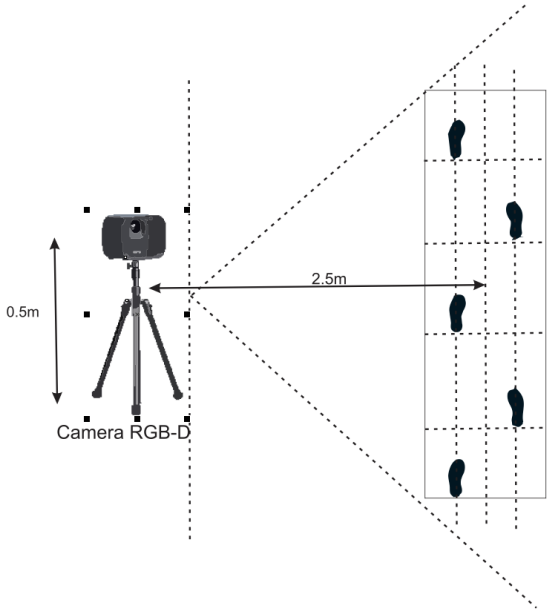


Fig. 2. Proposal for camera placement.

to ensure the validity and reliability of the obtained results. These databases have been collected with rigorous methodologies and have been widely used in the scientific literature, allowing direct comparison of results and validation of new analysis approaches. Furthermore, access to a wide variety of data allows for the consideration of multiple variables and a more comprehensive and detailed analysis of gait patterns. By basing the use of these databases, transparency and reproducibility of the research are guaranteed[19, 26, 21].

Using established databases in gait analysis is essential to ensure the validity and reliability of the obtained results. Table 2 shows a brief description of databases such as GaitRec[20], MotionSense[22], CMU Graphics Lab Database[15], OU-ISIR Gait Database[21], and CASIA Gait Database [32], allowing direct comparison of results and validation of new analysis approaches. Additionally, access to a wide variety of data allows for the consideration of multiple variables and a more comprehensive and detailed analysis of gait patterns. By basing the use of these databases, transparency and reproducibility of the research are guaranteed.

4 Multimodal Machine Learning

Multimodal machine learning refers to the ability of models to process and relate multiple modalities from sensors, images, text, audio, and video. This field focuses on building models that can jointly interpret multimodal signals,

Table 2. Data base gait parameters.

Data base	Description	Adquisition	Subjets
GaitRec[20]	Data for the evaluation of gait recognition algorithms. Includes multiple subjects and conditions	Inertial sensors, RGB-D cameras	744
MotionSense[22]	Data from mobile sensors for real-time analysis of IoT systems, including gait data captured by wearable devices.	Accelerometers and gyroscopes in smartphones	24
CMU Graphics Lab Motion Capture Database[15]	Motion capture database that includes various activities, including gait, collected with high-speed cameras.	Motion capture cameras	25
OU-ISIR Gait Database[21]	Treadmill gait dataset, captured under multiple conditions and with different subjects.	Video cameras	34
CASIA Gait Database[32]	Gait recognition database, which includes multiple views and recording conditions.	Video cameras	124

leveraging available data to enhance understanding and performance across various tasks. Characteristics of multimodal learning include integrating data from different sources, the ability to learn joint representations, and the capability to translate and align information across modalities [4]. Among the advantages of multimodal machine learning are increased robustness and accuracy in pattern recognition and classification, as well as improved capability to capture complex contexts and nuances that would be challenging to understand from a single modality [23].

The proposal presented in this work adopts a multimodal system for gait analysis that integrates data from multiple sensory sources such as RGB cameras, depth cameras, and inertial sensors, enabling precise and comprehensive three-dimensional motion capture. This integration enhances the analysis by providing combined kinematic and kinetic data, allowing for a deep understanding of gait and its disorders. It highlights the potential of multimodal systems to significantly improve the understanding and treatment of gait disorders.

4.1 Multimodal Gait Parameters

The identification of parameters is essential for understanding and evaluating specific aspects of human gait, considering that they come from the combination of inertial sensors and depth cameras. Table 3 groups the parameters that can

Table 3. Multimodal Machine Learning comparison.

Parameter	RGB-D Cameras	Inertial Sensors	Common References
Position and joint angles	Yes	No	[20][21][32]
Speed and acceleration of movement	Yes	Yes	[20][21][32][22]
Body segment trajectories	Yes	No	[20][21][32]
Distances and step lengths	Yes	No	[20][21][32]
Area of movement	Yes	No	[20][21][32]
Detection of joint points	Yes	No	[20][21][32]
Linear and angular acceleration	No	Yes	[22]
Angular velocity	No	Yes	[22]
Orientation and posture	No	Yes	[22]
Step frequency	No	Yes	[22]
Duration of gait phases	No	Yes	[22]
Variability in movement patterns	No	Yes	[22]

be identified by each of the technologies used, facilitating the analysis and understanding of gait movement.

When using inertial sensors to record accelerations and rotations during gait analysis, various features can be extracted that are crucial for understanding and evaluating human movement. These features not only provide a quantitative description of movement during gait but can also serve as inputs for machine learning algorithms aimed at identifying specific patterns, recognizing anomalies, or classifying different gait conditions. The appropriate selection of these features depends on the study or clinical application's objectives and the type of biomechanical analysis desired.

RGB and depth camera systems are known for their ability to capture three-dimensional data using structured light technology, making them useful for gait analysis and other human motion studies.

Table 3 provides a clear and concise comparison of the features recorded by RGB-D cameras and inertial sensors in gait analysis. By identifying the overlaps and differences in the parameters measured by both technologies, the selection of appropriate tools for specific human gait studies is facilitated. This comparison also highlights the complementarity of both technologies, suggesting that a multimodal integration can offer a more comprehensive and accurate view of gait analysis, improving the detection and treatment of pathologies.

4.2 Machine Learning Algorithms

Studying the algorithms used in gait analysis is crucial for several reasons. Firstly, different algorithms may offer varying levels of precision and efficiency,

Table 4. Multimodal Machine Learning Algorithms.

Reference	Algorithm	Features	Evaluation Metric (%)	Database
Perry, J., & Burnfield, J. M. (2010)	SVM	Temporal and spatial gait analysis	85%- accuracy	GaitRec
Kirtley, C. (2006)	Random Forest	Acceleration and gyroscope features	88%- accuracy	Daphnet
Umphred, D. A., et al. (2013)	CNN	Images and depth sequences	90%- specificity	Murooka
O'Sullivan, S. B., et al. (2019)	Decision Trees	Frequency and time parameters	82%- accuracy	GaitRec
Buczek Jr., F. L., et al. (Year)	LSTM	Temporal movement sequences	87%- perplexity-accuracy	Daphnet
Webster, J., & Murphy, D. (2018)	KNN	Joint angle analysis	80%- log loss	Murooka
Journal of Biomechanics	Naive Bayes	Kinematic parameters	83%- recall-accuracy	GaitRec
Journal of Biomechanical Engineering	AdaBoost	Combination of temporal and spatial features	89%-F1Score	Daphnet

allowing the selection of the most suitable algorithm for specific study or application needs. Additionally, some algorithms are better suited for integrating and analyzing data from multiple sources, such as inertial sensors and RGB-D cameras, optimizing multimodal analysis. Advanced algorithms, such as neural networks and machine learning models, can identify complex patterns in gait data that may not be detectable using traditional methods, which is essential for the diagnosis and treatment of gait pathologies. In clinical settings, choosing the correct algorithm can significantly improve the diagnosis, monitoring, and treatment of patients, providing more reliable and replicable results. Understanding the algorithms used also drives research and the development of new technologies and methods, contributing to the advancement of the field.

Table 4 provides a comparative overview of various studies employing machine learning algorithms for gait pathology detection, highlighting their importance in validation and methodology comparison. The cited references ensure the validity of the results, while the diversity of algorithms such as SVM, Random Forest, and CNN, demonstrates the breadth of applicable approaches. The utilized features, including temporal and spatial analysis, acceleration

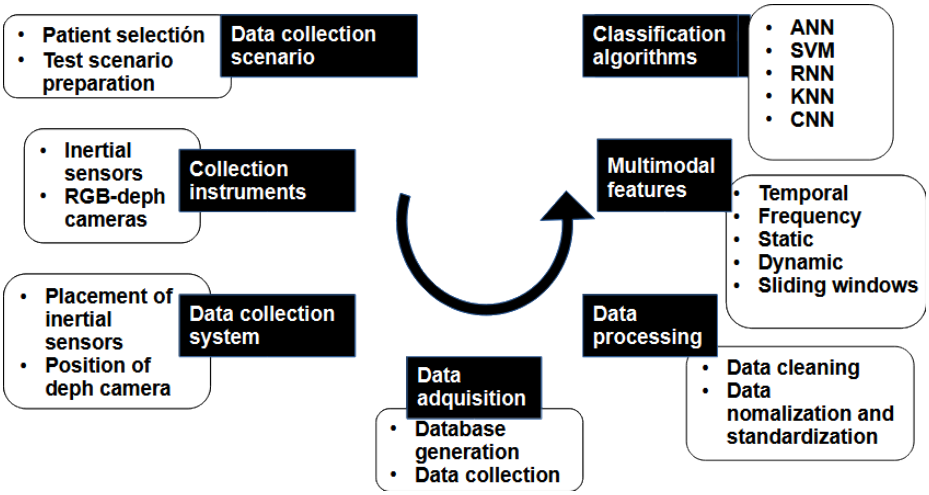


Fig. 3. Proposed architecture.

and gyroscope data, images, and depth sequences, are crucial for capturing relevant signals. The accuracy evaluation, ranging from 80% to 90%, allows for comparison of approach effectiveness. Additionally, databases like GaitRec, Daphnet, and Murooka ensure result validity and generalizability.

4.3 Proposed Approach

The proposed methodology for the multimodal integration of inertial sensors and RGB-D cameras is shown in Fig.3, where a detailed structure for gait analysis is presented. This approach utilizes technologies common to the previously reviewed works, supplemented with preprocessing techniques and machine learning algorithms. The multimodal approach allows for a deep understanding of gait patterns and facilitates the identification of pathologies in human locomotion.

The first step in the gait analysis methodology is to develop an appropriate testing environment. This environment must be controlled and standardized to ensure consistent and reproducible testing conditions. A sufficiently large and flat area should be selected to allow for natural walking, with specific distances marked for participants to walk. It's important to consider lighting and the absence of obstacles that could interfere with walking.

The selection of data collection instruments is crucial for obtaining accurate and useful measurements. In this methodology, inertial measurement units (IMUs) and RGB-D cameras will be used. IMUs are useful for measuring accelerations and rotations, while RGB-D cameras capture images and depth data, providing detailed information about body movement and position in space.

Proper sensor placement is essential for obtaining accurate data. IMUs should be placed at key points on the body such as ankles, knees, hips, and the lower back to capture limb and trunk movement. RGB-D cameras should be positioned around the testing environment to cover multiple angles and ensure that the entire gait sequence is captured without obstructions. Ideal placement is typically at mid-height and at the ends of the walking area to maximize coverage and data depth accuracy.

Once sensors are placed, data acquisition proceeds. Participants walk along the testing environment while IMUs and RGB-D cameras record their movements. It's important to conduct multiple trials for each participant to obtain a robust dataset and better represent natural variations in gait. Data should be properly stored and labeled to facilitate subsequent processing and analysis.

Data preprocessing is a critical step to ensure that the obtained data are of high quality and suitable for analysis. This process includes cleaning data to remove noise, synchronizing data between different sensors, and normalizing data to adjust for differences in measurement scale. Data can also be segmented into individual gait cycles to facilitate specific analysis of each phase of gait.

Once preprocessed, the data are analyzed to extract relevant features. Temporal features include parameters such as gait cycle duration and individual phases (stance and swing). Spatial features include measures such as step length, step width, and pelvic tilt. Frequency-domain features are obtained through spectral analysis, identifying dominant frequencies in motion signals that may be related to specific gait patterns or pathologies.

The final step is the application of multimodal classification algorithms to analyze the extracted features and classify gait patterns. Machine learning algorithms such as Support Vector Machines (SVM), Random Forest, and neural networks (e.g., LSTM for temporal data) are trained using features extracted from IMU and RGB-D data. These algorithms can identify and classify different types of gait, including normal and pathological patterns, enabling precise and detailed assessment of participants' gait.

5 Conclusions

In the study of human gait, technologies such as inertial sensors and Kinect cameras have been used, and various methodologies and applications in machine learning and biomechanics have been explored. It has been reviewed how inertial sensors capture acceleration and gyroscope data, which are crucial for analyzing parameters such as speed, cadence, and abnormal gait movement patterns. On the other hand, Kinect has proven useful for recording three-dimensional joint positions, enabling a detailed analysis of human movement kinematics and dynamics.

In terms of machine learning algorithms, the potential of Convolutional Neural Networks (CNNs) to process images captured by Kinect has been noted, as well as Recurrent Neural Networks (RNNs) for modeling the temporal

dynamics of inertial sensor data. Multimodal neural networks and other methods like Support Vector Machines (SVMs) have been considered to integrate and classify data from multiple sources.

Finally, it is envisaged how these technologies and methodologies can significantly contribute to medical diagnosis, rehabilitation, and diagnostic improvement, providing the capability to predict the onset of gait problems in individuals.

References

1. Alanazi, M.A., Alhazmi, A.K., Alsattam, O., Gnau, K., Brown, M., Thiel, S., Jackson, K., Chodavarapu, V.P.: Towards a low-cost solution for gait analysis using millimeter wave sensor and machine learning. Sensors 22(15), 5470 (2022)
2. Albert, J.A., Owolabi, V., Gebel, A., Brahms, C.M., Granacher, U., Arnrich, B.: Evaluation of the pose tracking performance of the azure kinect and kinect v2 for gait analysis in comparison with a gold standard: A pilot study. Sensors 20(18), 5104 (2020)
3. Arrieta Ramos, J.C., Herrera Velásquez, J.C., et al.: Diseño, implementación, configuración y puesta en marcha de un sistema integral de control de acceso utilizando equipos biométricos de reconocimiento facial con inteligencia artificial para la compañía liberty seguros
4. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence 41(2), 423–443 (2019)
5. Bijalwan, V., Semwal, V.B., Mandal, T.K.: Fusion of multi-sensor-based biomechanical gait analysis using vision and wearable sensor. IEEE Sensors Journal 21(13), 14213–14220 (2021)
6. Buisson Garcia, S., et al.: Segmentación de las fases de la marcha humana y predicción de sus eventos futuros mediante sensores inerciales y técnicas de aprendizaje máquina (2023)
7. Cai, S., Shao, M., Du, M., Bao, G., Fan, B.: A binocular-camera-assisted sensor-to-segment alignment method for inertial sensor-based human gait analysis. IEEE Sensors Journal 23(3), 2663–2671 (2023)
8. Caldas, R., Mundt, M., Potthast, W., Buarque de Lima Neto, F., Markert, B.: A systematic review of gait analysis methods based on inertial sensors and adaptive algorithms. Gait Posture 57, 204–210 (2017), https://www.sciencedirect.com/science/article/pii/S0966636217302424
9. Cedeño Barahona, R.A., Párraga Mendoza, J.G., Loayza Paredes, F., et al.: Diseño de estimulador con inteligencia artificial para el congelamiento de la marcha en la enfermedad de Parkinson. Ph.D. thesis, ESPOL. FIMCP (2022)
10. Chen, C., Jafari, R., Kehtarnavaz, N.: A review of depth and inertial sensor fusion for human action recognition. Multimedia Tools and Applications 76(3), 4405–4425 (2015)
11. Díez Díaz, F., Quirós, P., García Fernández, S., Pedrosa García, I., et al.: Detección temprana de enfermedades asociadas a la marcha mediante tecnología edge computing en entorno extraclínico (2022)
12. D'Antonio, E., Taborri, J., Mileti, I., Rossi, S., Patané, F.: Validation of a 3d markerless system for gait analysis based on openpose and two rgb webcams. IEEE Sensors Journal 21(15), 17064–17075 (2021)

13. Eskofier, B.M., Lee, S.I., Daneault, J.F., Golabchi, F.N., Ferreira-Carvalho, G., Vergara-Diaz, G.P., Sapienza, S., Costante, G.: Recent machine learning advancements in sensor-based mobility analysis: Deep learning for gait and fall risk assessment. Frontiers in Computer Science 1, 8 (2017)
14. Gabel, M., Gilad-Bachrach, R., Renshaw, E., Schuster, A.: Full body gait analysis with kinect. In: 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society. pp. 1964–1967. IEEE (2012)
15. Gross, R., Shi, J.: The cmu motion of body (mobo) database. In: Proceedings of the IEEE Workshop on Human Motion. pp. 136–141 (2001)
16. Guffanti, D., Brunete, A., Hernando, M.: Non-invasive multi-camera gait analysis system and its application to gender classification. IEEE Access 8, 95734–95746 (2020)
17. Ionescu, C., Moga, S.: Gait recognition based on multiple projections using machine learning algorithms. International Journal of Computers Communications & Control 10(1), 29–38 (2015)
18. van Kersbergen, J., Otte, K., de Vries, N.M., Bloem, B.R., Röhling, H.M., Mansow-Model, S., van der Kolk, N.M., Overeem, S., Zinger, S., van Gilst, M.M.: Camera-based objective measures of parkinson's disease gait features. BMC research notes 14, 1–6 (2021)
19. Kirtley, C.: Clinical Gait Analysis: Theory and Practice. Elsevier Health Sciences, Edinburgh (2006)
20. Künzner, L.: Gaitrec: A large-scale ground-truth resource for evaluating gait recognition algorithms. arXiv preprint arXiv:1703.01002 (2017)
21. Makihara, Y., Muramatsu, D., Echigo, T., Yagi, Y.: The ou-isir gait database comprising the treadmill dataset. IPSJ Transactions on Computer Vision and Applications 4, 53–62 (2012)
22. Malekzadeh, M., Sain, S., Welsh, R., Barker, S., Yates, D.C., Bagci, I.E., Dasmahapatra, S., Gopalakrishnan, A., Rahman, M.S.I., Crowcroft, J.: Mobile sensor data analytics for real-time control of iot systems. In: Proceedings of the 20th International Workshop on Mobile Computing Systems and Applications. pp. 69–74 (2019)
23. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11). pp. 689–696 (2011)
24. Palermo, M., Lopes, J.M., André, J., Matias, A.C., Cerqueira, J., Santos, C.P.: A multi-camera and multimodal dataset for posture and gait analysis. Scientific data 9(1), 603 (2022)
25. Panwar, M., Gupta, P.: A review of gait recognition techniques using kinect sensor. Multimedia Tools and Applications 78, 16891–16921 (2019)
26. Perry, J., Burnfield, J.M.: Gait Analysis: Normal and Pathological Function. SLACK Incorporated, Thorofare, NJ, 2nd edn. (2010)
27. Rueda, F.M., Tejada, M.C.: La marcha humana: biomecánica, evaluación y patología. Panamericana (2020)
28. Sprager, S., Juric, M.B.: Inertial sensor-based gait recognition: A review. Sensors 15(9), 22089–22127 (2015), https://www.mdpi.com/1424-8220/15/9/22089
29. Springer, S., Yogev Seligmann, G.: Validity of the kinect for gait assessment: A focused review. Sensors 16(2), 194 (2016)
30. Wang, L., Tan, T., Ning, H., Hu, W.: Silhouette analysis-based gait recognition for human identification. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(12), 1505–1518 (2003)

31. Yamamoto, M., Shimatani, K., Ishige, Y., Takemura, H.: Verification of gait analysis method fusing camera-based pose estimation and an imu sensor in various gait conditions. Scientific reports 12(1), 17719 (2022)
32. Yu, S., Tan, T., Huang, K.: Casia gait database: A free resource for gait recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence 28(7), 1333–1335 (2006)
33. Zhang, Z., Han, G., Wu, H., Yin, Y.: Deep multimodal representation learning: A survey. IEEE Transactions on Cognitive and Developmental Systems (2021)

ISSN 1870-4069

Emotion Analysis in Children's Drawings Using Machine Learning Techniques

Ricardo Ramos-Aguilar, Daniel Sánchez-Ruiz, Karla Rivera-Lima, Estefani Jaramillo-Nava, Natali Meza-Barranco

Instituto Politécnico Nacional,
Unidad Interdisciplinaria de Ingeniería Campus Tlaxcala,
Mexico

{rramosa,dsanchezro}@ipn.mx,
{kriveral2100,ejaramillon2200,nmezab2100@}@alumno.ipn.mx

Abstract. Emotions are a set of chemical and neural responses that regulate the body to act in various situations. They can be triggered automatically and are fundamental to human life. From an early age, emotions play an important role in the personal development of human beings, as they impact emotional intelligence, social and relational development, mental health, decision-making, academic learning, and overall well-being. This work presents a proposal for the non-invasive detection of emotions in children through the analysis of their drawings, using machine learning techniques, manual feature extraction and deep learning methods. A methodology and a proposal are presented in this work, with all their stages listed in general, following the specifics of each stage. Finally, the conclusions and future work are described.

Keywords: Emotion recognition, children emotion recognition, pattern recognition.

1 Introduction

Emotions are psychological reactions in human beings that occur as a way of adaptation when they face different situations, places, people, objects, and more. These are important because they control impulses, express sentiments, manage feelings, behaviors, and more. Understanding emotions has become essential for the daily functioning of humans, as the acquisition of emotions and experiences is crucial for interpersonal communication in social environments. An emotion influences on the thoughts and interfere in decisions, actions, memories and perceptions [26]. Human emotions can be recognized through studies that analyzed facial expressions, tone of voice, posture, eye movements, appearance, and verbal cues. These analyses are conducted by individuals who understand human behavior, such as psychologists.

Currently, computational techniques allow us to recognize emotions through speech, text, facial images, and electroencephalograms. Facial expressions are the most significant factor for recognizing human emotions and act as a key

element in understanding and perceiving the mental state of any individual, thereby guiding subsequent actions. This has become a key reason in numerous research works, where several hypotheses, experiments, and practical studies are still being carried out [15].

On the other hand, speech is a natural form of communication between humans. It provides information about thoughts, feelings, moods, and the context of the speaker's communication. The subtleties of emphasis, tone, phrasing, variations in speed and continuity of expression and the accompanying physical gestures convey something of the inner life of impulse and feeling [12].

Functional neuroimaging techniques such as electroencephalography (EEG), functional magnetic resonance imaging (fMRI), or positron emission tomography(PET) can be used. Although EEG has poor spatial resolution and requires many electrodes placed in various locations on the head, it provides excellent temporal information, allowing researchers to study phase changes in response to emotional stimuli. Additionally, EEG is non-invasive, fast, and cost-effective, making it the preferred method for studying brain responses to emotional stimuli [3].

Emotion recognition has potential in the healthcare sector as it can help physicians understand the emotional states of their patients, enabling the delivery of individualized and personalized medical services. Emotion recognition also has the ability to enhance intelligent customer service systems by detecting users' emotional needs and offering more personalized services. Furthermore, emotions have a substantial influence on human cognitive processes, including perception, attention, learning, memory, reasoning, and problem-solving. Emotion has a particularly strong impact on attention, especially modulating attentional selectivity and motivating action and behavior. This attentional and executive control is closely linked to learning processes, as intrinsically limited attentional capacities focus better on relevant information. Emotion also facilitates encoding and aids in the efficient retrieval of information [27], however, analyzing emotions is not the same for adults and infants, and the applications for both may differ. Nevertheless, the importance of understanding and/or analyzing emotions is equally significant at any age.

Analyzing emotions in children is crucial as it helps to recognize their emotional development and emotional intelligence, which can aid in regulating their emotions. Understanding emotions facilitates empathy and social skills, including interpersonal relationships and conflict resolution. Additionally, the ability to express and process emotions contributes to mental well-being, education, and learning. [13]. From the various approaches to identifying emotions, the interest in analyzing emotions in children arises.

Child abuse and neglect are critical issues worldwide, resulting in harm to the child's health, development, and dignity. It can be classified into four categories: physical abuse, psychological abuse, sexual abuse, and neglect. The long-term effects include diminished abilities to trust, form intimate relationships, and engage in healthy behavior, as well as a range of ongoing mental and physical

health issues. Additionally, the experience of abuse affects children's social cognition, particularly their ability to understand others [18].

Although disclosing child abuse can put an end to it, many survivors remain silent due to reasons such as fear and shame, difficulty in expressing their experience, or being silenced by the perpetrator or their supporters. Additionally, the impact of abusive experiences on the brain often results in dissociation and an inability to recall the traumatic events. These factors present significant challenges for investigators and practitioners in developing techniques to encourage victims to disclose their experiences [14].

The self-figure drawing, developed by Lev-Wiesel from the draw-a-person test, was designed to identify important drawing indicators in human figure drawings of individuals with traumatic experiences, as it can be used to assess specific psychological and emotional traits [14]. As Duque [8] states, drawing is considered a fundamental tool for emotional expression in childhood. The act of drawing provides the opportunity to make thoughts and emotions tangible. Through lines, colors, and shapes, feelings and internal states can be expressed creatively and visually. When children draw, they are capturing their emotions on paper in a unique and personal way. In this way, adults can use such drawings to try to better understand children's thoughts, while the children feel more understood and validated.

This work presents a proposal to recognize emotions through family drawings, which currently are used by educators, psychologists, and individuals studying human behavior, often involving a questionnaire to create an individual report. Children drawing analysis via machine learning algorithms is proposed with the aim of serving as a tool for parents and professionals working in the field of emotions. This procedure is non-invasive and could complement current child examinations.

The document is divided into the following sections: first, a background of useful terms for understanding is presented; second, the related works are presented; third, research method to develop the proposal is shown; fourth, the proposal to solve the problem is described; finally, the conclusions are provided.

2 Background

2.1 Emotions

Emotions are a set of chemical and neural responses, as well as physiological reactions that adapt or adjust the organism to face a specific phenomenon. They are useful for survival and arise automatically and spontaneously due to internal or external stimuli, and they can be modulated to some extent voluntarily or through external stimuli [25]. The brain is primarily responsible for processing basic emotions, mainly the limbic system and the prefrontal cortex, and it consists of five factors: motor expression, cognitive components, neurophysiological, motivational, and subjective experience [27].

In recent years, a growing number of studies have attempted to investigate the structure of affect. Most of these studies agree that affective experience

has two dominant dimensions, namely, positive affect (PA) and negative affect (NA). In Bradburn's pioneering work, PA and NA have been described as two independent unipolar dimensions of affect, encompassing all affective states with a positive valence (joy, enthusiasm, love, etc.) or a negative valence (anger, fear, anxiety, etc.). The two dimensions of affect (PA and NA) have been crucial in the conceptual differentiation between depressive and anxiety disorders. Furthermore, PA and NA have also been strongly associated with the personality dimensions of Extraversion and Neuroticism, respectively [7]. Therefore, the analysis and recognition of emotions have commonly been studied by specialists such as psychologists, mostly analyzing the complete behavior of human beings, such as facial expressions, body language, tone of voice, among other things, using interviews or data collection instruments in a direct and personal manner where human experience plays a central role in the analysis.

2.2 Emotions Categories

The discrete emotions theory is based on six categories of basic emotions: sadness, happiness, fear, anger, disgust, and surprise. These innate and culturally independent emotions are experienced over a short period of time. Other emotions are derived from combinations of these basic ones. Most existing emotion recognition systems (ERS) focus on these basic emotional categories. However, these discrete categories of emotions are not capable of defining some of the complex emotional states observed in daily communication [2]. Each emotional category is characterized by unique internal experiences, external expressions, and physiological patterns. Basic emotions must possess the following attributes: (1) Emotions originate from innate instincts; (2) Various individuals manifest the same emotion in response to the same circumstances; (3) People tend to express basic emotions in a similar manner; (4) The physiological patterns of different individuals are consistent when experiencing basic emotions. Additionally, some studies incorporate the classification of data into distinct categories, such as positive (i.e., expressing a positive feeling), negative (i.e., indicating a negative feeling), or neutral (i.e., lacking affective impact) [9].

2.3 Techniques

Some methods for investigating emotions primarily focus on physiological aspects, offering quantifiable data on the body's response to emotional stimuli, such as EEG, fMRI, ECG, and skin conductance. The study of emotions can be conducted consciously, using subjective methods and instruments, or unconsciously, by using objective methods and instruments. Objective measures should be used if the goal is to assess the most unconscious level of an individual, which can be achieved by using psychophysiological measures such as electrodermal activity (EDA), heart rate (HR), or electroencephalography. These allow for the collection of physiological and biometric data, providing essential information about how a person feels, even if they are not consciously

aware of it. However, they also present some drawbacks, primarily because they are intrusive and noisy [20].

In turn, subjective measures can be used if the goal is to evaluate the emotional experience from the individual's subjective point of view. This includes established user scales, such as the "Visual Analog Scale" (VAS) or the "Self-Assessment Manikin" (SAM), interviews, thinking aloud, and questionnaires (e.g., the "Check-All-That-Apply" procedure, the "Positive and Negative Affect Schedule", or the "State-Trait Anxiety Inventory"). These measures can be pictorial, like the SAM, which presents a representative drawing of the human figure for respondents to express their emotions.

In various areas of engineering with different applications, attempts have been made to analyze emotions. From a computational perspective, human-machine interaction requires understanding user emotions for specific tasks, such as adapting applications as needed. To achieve this, a series of steps are followed, such as acquiring data, preprocessing, analyzing, and applying. However, achieving this requires various sources, such as those mentioned below.

Facial expressions, behavior, voice, text, and physiological signals can all be used to identify human emotions. Although the effectiveness and objectivity of these methods are still under investigation, some studies suggest that physiological cues may provide more objective measures of emotional states in certain contexts than the other approaches. Emotion recognition (ER) involves studying human emotions and developing tools or systems that can track, analyze, understand, and respond to these emotions. A camera can record body postures, gestures, and facial expressions, while a microphone captures speech and sounds. Additionally, sensors are used to monitor physiological data such as heartbeat, skin temperature, blood pressure, pulse rates, and skin's galvanic resistance to detect emotional states. ER requires feature extraction, relevant pattern discovery, and data analysis [9].

2.4 Limitations

Some biases and stereotypes affect the accuracy of diagnoses, and often systems reflect more of the ideas of those who created them than a true classification of the characteristics of the people they aim to help. Cultural context can influence what is considered "healthy" or "pathological," which can lead to misdiagnoses in people from different cultures [11].

From one perspective, it should be easy to determine if someone is experiencing a particular emotion. However, Mauss suggests that measuring a person's emotional state is one of the most complex problems in affective science [21].

3 Related Work

This work presents a proposal to analyze emotions in children's drawings, and in this section, relevant works on machine learning and image-based approaches are discussed.

A master's thesis developed by Brinkman [6] presents a machine learning approach for emotion classification in abstract expressionist paintings. The research explores how abstract shapes, expressive colors, and textures in abstract expressionist paintings evoke emotions in human observers. The study aims to answer whether it is possible to predict people's emotional reactions when viewing these artworks. To this end, an emotion corpus was collected, including ratings on the dimensions of arousal, dominance, and valence, and image features from a set of paintings were analyzed. The machine learning algorithms used correctly classified emotions in the arousal dimension at 73.33%, while the accuracy was lower in the dominance 46.67% and valence 73.33% dimensions.

In 2017, Moetesum et al. developed a study to detect facial expressions in human figure drawings, focusing specifically on the face [22]. A computerized system was proposed to analyze hand-drawn facial images to extract expressions from the image. The subject draws a human face, which is then input into the system, where the image is binarized and segmented into different facial components. Features based on Local Binary Patterns (LBP), Gray Level Co-occurrence Matrices (GLCM), and Histograms of Oriented Gradients (HOG) computed from the facial components are used to train an SVM classifier to distinguish between four expression classes: 'happy,' 'sad,' 'angry,' and 'neutral.'

In 2022, an emotion recognition application was developed by Nasva et al. [4], based on artificial intelligence (AI) and called the Emotion Sensing Recognition App (ESRA) to help parents and teachers understand children's emotions by analyzing their drawings. A total of 102 drawings were collected from a local school in Doha and 521 drawings from Google and Instagram. The deep learning model was trained using the Fastai library and ResNet. The model classifies the drawings into positive or negative emotions, with an accuracy ranging from 55% to 79% in the four experiments.

A study developed in 2024 by Khan et al. [16] to analyze emotions specifically uses behavioral metrics such as drawing and handwriting to determine a person's emotional state, recognizing these actions as physical functions that integrate motor and cognitive processes. The study proposes an attention-based transformer model as an innovative approach to identify emotions from handwriting and drawing samples, thereby advancing the capabilities of ER into the domains of fine motor skills and artistic expression. The proposed method achieved cutting-edge results with 92.64% accuracy on the benchmark dataset known as EMOTHAW (Emotion Recognition via Handwriting and Drawing).

The art of drawing is a substantial conduit for delving into the rich complexities of human emotions, facilitating unique insights into non-verbal emotional expressions. In the work developed by Weng [28], it is mentioned that drawings function as a means to reveal emotional expressions. Moreover, obtaining family drawings presents a non-invasive instrument compared to other methods and represents a projective technique widely used by clinical and developmental psychologists to access a child's inner world [24].

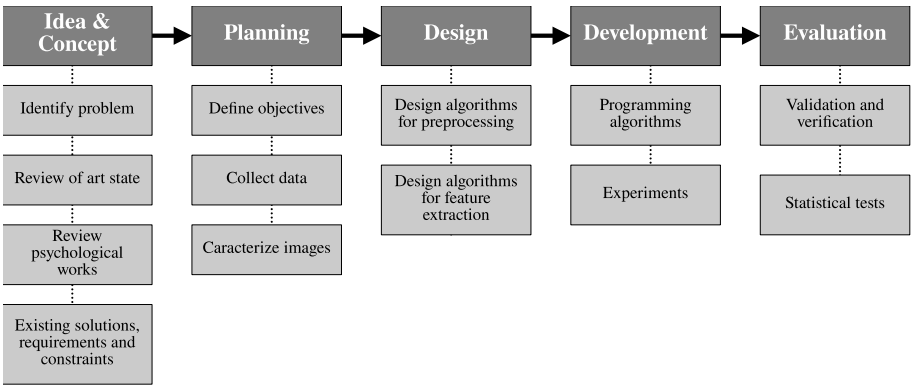


Fig. 1. Methodology of work for the development of research proposal.

As observed in the different studies, it currently presents a challenge due to the lack of information. However, the advancements made have shown that drawings can give us a perspective on emotions. Most studies attempt to evaluate basic emotions, some grouping them into positive or negative categories, or based on the dimension they represent. Among the studies found, most present approaches based on manual feature extraction, such as color, shape, texture, time, frequency, and statistics. Only Khan's work [16] presents a methodology implementing transformers, similar to what is intended for the proposed methodology, and they also use an available database (EMOTHAW) consisting of images and text that represent adult emotions. Generally, working with children's datasets is challenging due to data collection issues; despite being a useful method for detecting emotions through drawings, it may have certain limitations because drawing skills depend on gender and age.

4 Research Methodology

This section describes the steps shown in Fig. 1, where the research stages for the development of the proposal are presented.

Idea and Concept: In this stage, the problem is defined, and identified through interviews in schools about various issues, with a particular focus on the emotional state of children. This is difficult to identify because they do not express it concretely, and it is challenging to approach them without involving their parents. After identifying the problem, related works were reviewed, specifically on the automatic identification of emotions from drawings using machine learning techniques, as well as studies in the psychological field. Finally, existing solutions were reviewed, including their requirements, limitations, advantages, disadvantages, and how current methods address the problem from the proposed approach.

Planning: In this stage, the general and specific objectives are defined. The main objective is to recognize basic emotions in grayscale drawings using machine

learning techniques, with two options presented. A dataset of children's drawings will be created with quick direct labeling by the children themselves; however, the drawings will also be analyzed by experts to ensure the reliability of the emotions in the drawings. Finally, some of the images will be characterized to design automatic description algorithms.

Design: Based on the proposed objective, there are two approaches to achieve it. The first is through emotion analysis using manual feature extraction methods, and the second option involves the implementation of deep learning methods, in particular via a transformer. To achieve this, an algorithm will be developed for the preprocessing of images, including enhancement and scaling. Subsequently, feature extraction techniques will be implemented.

Development: This stage involves programming the designed algorithms, as well as conducting various experiments, such as evaluating different classification algorithms, their parameters, and data partitions (i.e., all classes or the division of positive and negative emotions).

Evaluation: An evaluation will be conducted using training, test, and validation sets, as well as the implementation of cross-validation techniques, area under the curve (AUC), and confusion matrices. Statistical metrics such as accuracy, precision, F-score, and others will be obtained.

5 Proposal

Emotion recognition has garnered growing interest from researchers across various fields. Human emotions can be identified through facial expressions, speech, behavior (such as gestures and posture), or physiological signals. However, the first three methods can be unreliable because individuals may unintentionally or intentionally hide their true emotions, a phenomenon known as social masking [29].

Therefore the use of physiological signals have become more attractive because it offers a more objective and reliable approach to emotion recognition. Compared to peripheral neurophysiological signals, electroencephalogram (EEG) signals are more sensitive to changes in affective states and can respond in real-time, providing valuable insights into emotional states. As a result, numerous EEG-based emotion recognition techniques have been developed in recent years [29]. However, these types of methods are intrusive and uncomfortable for many people, and, the commercialization and development of applications for regular users are still limited.

Since the approaches described can either conceal a person's true emotions or rely on invasive devices with limited accessibility for the general population, the alternative of using drawings, where each child expresses their ideas, is explored. This makes it an attractive, non-invasive, and personal option. As mentioned, datasets [1, 19] have been developed that utilize drawings made by children, leading to various studies [5, 10, 17, 23]. However, the disadvantage of these datasets and studies is that the drawings are part of psychological tests

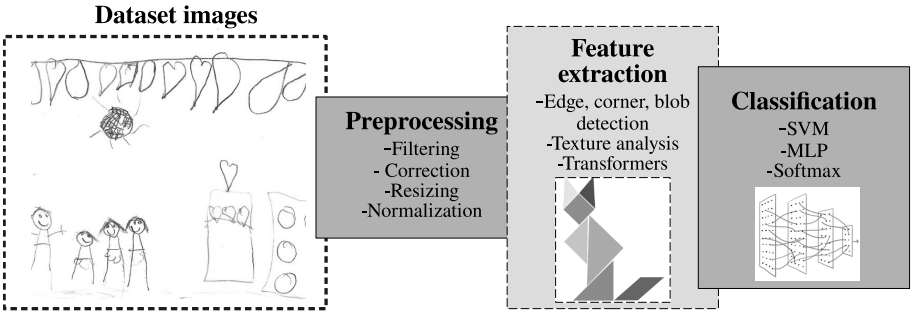


Fig. 2. Design of the proposal.

where the child is directed on what to draw, meaning that feelings and emotions are not allowed to be expressed naturally.

Taking into account all of this, the proposed approach for analyzing emotions in children based in drawings is depicted in Fig. 2. The first stage is data acquisition. At this point, the protocol for acquiring a new dataset will be established. The main difference between this dataset and existing ones is that the drawings will be requested freely and as part of a routine activity. Subsequently, with the expert knowledge of a mental health professional, the data will be labeled, and the emotions present will be identified. Doing it in this way ensures that the children are not directed, so unlike other studies, the data will be grouped later based on the identified emotions.

Since the drawings are going to be done on paper, not digitally, they need to be converted into a digital format. This process often results in an inherent loss of information. Additionally, it's common that the materials used by children to create the drawings to vary in their characteristics (color, graphite quality, etc.), so the graphical representations may not always have the same quality in terms of the information contained. For these reasons, a preprocessing stage is necessary, during which various filters and operations will be applied to maximize and standardize all the data in the dataset. These issues can be observed in the Fig. 3, which shows a couple of examples of the data captured in the ongoing work.

Feature extraction is the next stage and it can be categorized into shallow and deep features. Shallow features are hand-crafted and derived from various analysis domains, such as spatial-domain, time-domain or frequency-domain. Higher-dimensional features are often reduced using techniques like principal component analysis (PCA) or empirical cumulative distribution functions. Some studies [5,17,19,23] that have utilized drawings have focused on features related to the stroke, analyzing it in terms of its temporal aspects, such as the time a stroke remains uniform or is sustained, as well as through ductus (ductus is the way how strokes are drawn including stroke order, direction and speed) analysis. These descriptors will be evaluated; however, it is also proposed to analyze the spatial domain of the data through shape or texture using statistical descriptors.

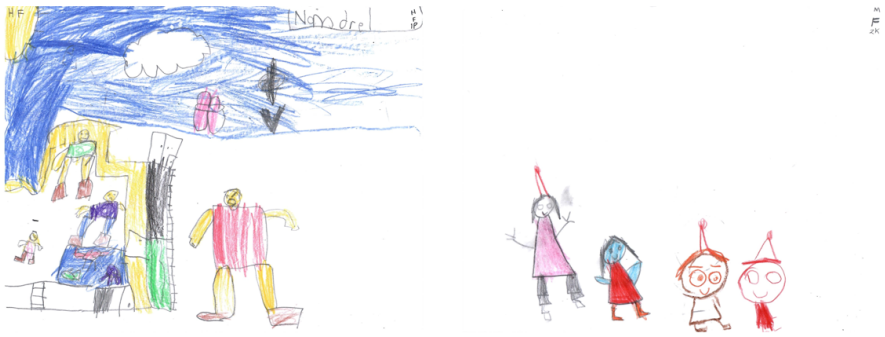


Fig. 3. Examples of acquired data and their characteristics.

To address the challenges of extracting effective and robust features, many researchers have turned to deep learning (DL) approaches. DL reduces the need for manually extracting features for machine learning models, as it can automatically learn a hierarchical feature representation [29]. This approach eliminates the need for data preprocessing and feature space reconstruction typically required in a standard machine learning pipeline. Feature extraction can be performed through convolutional neural networks or transformers, which is the second approach proposed for obtaining descriptors from the dataset.

After any type of feature extraction method, a classification stage will be carry out. Several machine learning methods will be employed, e. g. Support Vector Machine, Multilayer Perceptron or Random Forest. An optimization stage it must need in this part of the proposal in order to obtained the best results for every method. For evaluation, accuracy and precision metrics will be used, applying cross-validation and partitioning the dataset for training, testing, and validation.

6 Conclusions

Emotions are an inherent part of all humans that help regulate physical stimuli, express feelings, and communicate personal ideas and sensations. Emotions can be expressed in multiple ways, through physical reactions, texts, drawings, or neurologically. Recognizing them helps in understanding what a person is thinking at a specific moment. Unlike adults, children do not always have the confidence or vocabulary to express themselves, therefore, analyzing drawings developed by them could facilitate the analysis of their emotions.

Drawings are a pictorial tool for the manifestation of emotions that has been researched through theoretical approaches; however, their analysis and recognition through technological developments have not been explored in depth. Analyzing emotions through children's drawings is a useful and accessible

alternative for understanding the emotional behavior of children. However, computationally it presents a significant challenge, as the development of sketches, lines, drawings, and other elements that a drawing may contain varies greatly among children due to their skills and ages, leading to marked differences between individuals when drawing elements or objects.

This proposal presented a methodology for analyzing drawings made by children in a non-invasive way using machine learning techniques. The first point emphasized is the importance of collecting a dataset validated by mental health experts. Additionally, the proposal aims to use drawings with colors and other types of scenarios that will represent an even greater challenge due to all the differences that images may present based on the subjects' personalities for the study, so a preprocessing stage it must be performed. The study of related works to define feature extraction methods and recognition methods is also fundamental. Finally, defining appropriate metrics for evaluation and comparing with related works are necessary steps to carry out.

Future work involves completing the data collection and labeling with the guidance of mental health experts. Subsequently, an analysis of the data needs to be conducted and the most appropriate preprocessing operations defined. Once this is done, methods and techniques for feature extraction can be defined and implemented. Finally, recognition methods should be implemented, experiments designed, and the results evaluated.

Acknowledgments. We extend our sincere gratitude to the Division of Operation and Research Promotion of the Instituto Politécnico Nacional for their invaluable support for the project registered 20242035.

References

1. Ahmadsaraei, M.F., Bastanfard, A., Amini, A.: Child psychological drawing pattern detection on obget dataset, a case study on accuracy based on myolo v5 and mresnet 50. Multimedia Tools and Applications 83(13), 39283–39313 (2024)
2. Akçay, M.B., Oğuz, K.: Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Communication 116, 56–76 (2020), https://www.sciencedirect.com/science/article/pii/S0167639319302262
3. Alarcão, S.M., Fonseca, M.J.: Emotions recognition using eeg signals: A survey. IEEE Transactions on Affective Computing 10(3), 374–393 (2019)
4. Ali, N., Abd-Alrazaq, A., Shah, Z., Alajlani, M., Alam, T., Househ, M.: Artificial intelligence-based mobile application for sensing children emotion through drawings. In: Studies in Health Technology and Informatics. Studies in health technology and informatics, IOS Press (Jun 2022)
5. Ayzeren, Y.B., Erbilek, M., Çelebi, E.: Emotional state prediction from online handwriting and signature biometrics. IEEE Access 7, 164759–164774 (2019)
6. Brinkman, E.: Recognition of the emotion in abstract expressionistic paintings using a machine learning approach (January 2008), http://essay.utwente.nl/863/
7. Díaz-García, A., González-Robles, A., Mor, S., Mira, A., Quero, S., García-Palacios, A., Baños, R.M., Botella, C.: Positive and negative affect schedule

(PANAS): psychometric properties of the online spanish version in a clinical sample with emotional disorders. BMC Psychiatry 20(1), 56 (Feb 2020)
8. Duque Martín, P., Vargas Soria, P.: El dibujo infantil, una gran ventana de expresión (2017)
9. Ezzameli, K., Mahersia, H.: Emotion recognition from unimodal to multimodal analysis: A review. Information Fusion 99, 101847 (2023), https://www.sciencedirect.com/science/article/pii/S156625352300163X
10. Fathi Ahmadsaraei, M., Bastanfard, A., Amini, A.: Obgess: Automating original bender gestalt test based on one stage deep learning. International Journal of Computational Intelligence Systems 16(1), 178 (2023)
11. Gallagher, M., Lopez, S.: Positive psychological assessment. Washington, DC: american Psychological association (2019)
12. George, S.M., Muhamed Ilyas, P.: A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise. Neurocomputing 568, 127015 (2024), https://www.sciencedirect.com/science/article/pii/S0925231223011384
13. Ioannidou, F., Konstantikaki, V.: Empathy and emotional intelligence: What is it really about. International Journal of Caring Sciences 1 (01 2008)
14. Jaroenkajornkij, N., Lev-Wiesel, R., Binson, B.: Use of self-figure drawing as an assessment tool for child abuse: differentiating between sexual, physical, and emotional abuse. Children 9(6), 868 (2022)
15. Kassam, K.S., Markey, A.R., Cherkassky, V.L., Loewenstein, G., Just, M.A.: Identifying emotions on the basis of neural activation. PLOS ONE 8(6), 1–12 (06 2013), https://doi.org/10.1371/journal.pone.0066032
16. Khan, Z.A., Xia, Y., Aurangzeb, K., Khaliq, F., Alam, M., Khan, J.A., Anwar, M.S.: Emotion detection from handwriting and drawing samples using an attention-based transformer model. PeerJ Comput. Sci. 10, e1887 (Mar 2024)
17. Khan, Z.A., Xia, Y., Aurangzeb, K., Khaliq, F., Alam, M., Khan, J.A., Anwar, M.S.: Emotion detection from handwriting and drawing samples using an attention-based transformer model. PeerJ Computer Science 10, e1887 (2024)
18. Koizumi, M., Takagishi, H.: The relationship between child maltreatment and emotion recognition. PLOS ONE 9(1), 1–4 (01 2014), https://doi.org/10.1371/journal.pone.0086093
19. Likforman-Sulem, L., Esposito, A., Faundez-Zanuy, M., Clémençon, S., Cordasco, G.: Emothaw: A novel database for emotional state recognition from handwriting and drawing. IEEE Transactions on Human-Machine Systems 47(2), 273–284 (2017)
20. Magalhães, M., Coelho, A., Melo, M., Bessa, M.: Measuring users' emotional responses in multisensory virtual reality: a systematic literature review 83(14), 43377–43417 (Apr 2024)
21. Mauss, I.B., Robinson, M.D.: Measures of emotion: A review. Cogn. Emot. 23(2), 209–237 (Feb 2009)
22. Moetesum, M., Aslam, T., Saeed, H., Siddiqi, I., Masroor, U.: Sketch-based facial expression recognition for human figure drawing psychological test. In: 2017 International Conference on Frontiers of Information Technology (FIT). pp. 258–263 (2017)
23. Nolazco-Flores, J.A., Faundez-Zanuy, M., Velázquez-Flores, O.A., Cordasco, G., Esposito, A.: Emotional state recognition performance improvement on a handwriting and drawing task. IEEE Access 9, 28496–28504 (2021)
24. Pace, C.S., Muzi, S., Vizzino, F.: Family drawing for assessing attachment in children: Weaknesses and strengths. Frontiers in Psychology 13 (2022)

25. Palomero-Gallagher, N., Amunts, K.: A short review on emotion processing: a lateralized network of neuronal networks. Brain Struct. Funct. 227(2), 673–684 (Mar 2022)
26. Rathod, M., Dalvi, C., Kaur, K., Patil, S., Gite, S., Kamat, P., Kotecha, K., Abraham, A., Gabralla, L.A.: Kids' emotion recognition using various deep-learning models with explainable ai. Sensors 22(20) (2022)
27. Tyng, C.M., Amin, H.U., Saad, M.N.M., Malik, A.S.: The influences of emotion on learning and memory. Frontiers in Psychology 8 (2017)
28. Weng, H.C., Huang, L.Y., Imcha, L., Huang, P.C., Yang, C.T., Lin, C.Y., Li, P.H.: Drawing as a window to emotion with insights from tech-transformed participant images. Scientific Reports 14(1), 11571 (May 2024)
29. Zhang, J., Yin, Z., Chen, P., Nichele, S.: Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. Information Fusion 59, 103–126 (2020)

Electronic edition
Available online: http://www.rcs.cic.ipn.mx

http://rcs.cic.ipn.mx

Centro de Investigación
en Computación