XLIV Jornadas de Automática 2023

https://jautomatica.es

Estimación de grupos conversacionales empleando cámaras 3D y aprendizaje automático para su aplicación a robótica social

Delgado, D.[a,*], Praena, J.A.[a], Caballero, F.[b], Gómez, R.[c], Merino, L.[a]

[a] *Escuela Politécnica Superior, Universidad Pablo de Olavide, Crta. Utrera km 1, 41013, Sevilla, España.*
[b] *Escuela Técnica Superior de Ingeniería, Universidad de Sevilla, Sevilla, España.*
[c] *Honda Research Institute Japan, Tokio, Japón.*

To cite this article: Delgado, D., Praena, J.A.,Caballero, F., Gomez, R., Merino, L., 2023. Estimating groups of interacting persons using 3D cameras and machine learning for the application to social robotics. XLIV Jornadas de Automática, 575-580. https://doi.org/10.17979/spudc.9788497498609.575

Resumen

Este estudio se sitúa en el ámbito de la robótica social, explorando la detección de ”F-Formations”, estructuras espaciales que se producen en la interacción grupal humana. El artículo presenta un sistema de percepción de grupos conversacionales. El sistema puede emplearse para mejorar las habilidades de robots para interpretar contextos sociales. Basándose en obtención de datos de esqueletos de los participantes a través de una cámara Azure Kinect, en el artículo se presenta un estudio de la capacidad de varios algoritmos de aprendizaje automático para la estimación de F-Formations. Como resultado, se ha desarrollado un módulo ROS capaz de reconocer grupos de personas.

Palabras clave: Tecnología de robótica, Percepción y detección, Robótica inteligente

Estimating conversational groups using 3D cameras and machine learning for its application to social robotics

Abstract

This study is situated within the domain of social robotics, specifically investigating the detection of F-Formations, spatial structures that emerge during human group interaction. The article presents perception system to detect conversational groups. The system can be used to improve the abilities of robots to interpret social contexts. The article presents a study of the capacity of various machine learning algorithms for the estimation of F-Formations from the 3D skeletons of the participants as obtained by a 3D Azure Kinect camera. As a result, a ROS module capable of recognizing groups of people has been developed.

Keywords: Robotics technology, Perception and sensing, Intelligent robotics

1. Introducción

La robótica social es un campo emergente que se centra en el diseño de robots capaces de interactuar de manera efectiva con humanos en contextos sociales. Dichas interacciones suelen involucrar grupos de personas, por lo que un aspecto fundamental de estas es la identificación de las ”F-Formations”(Kendon, 1990), agrupaciones de personas que interactúan entre sí. La detección precisa de las F-Formations en entornos dinámicos sigue siendo un desafío.

Para llevar a cabo este estudio sobre F-Formations, se ha utilizado a Haru (Gomez et al., 2020) ,un robot social cuyo diseño integral tiene como objetivo la expresividad robótica (ver Fig. 1). Su diseño no sólo considera su físico, sino también su comportamiento y cómo se comunica y responde a los estímulos del ambiente. El objetivo es que Haru es el de ser un robot de sobremesa al que los humanos puedan acudir a para relacionarse y realizar actividades junto a él.

*Autor para correspondencia: diedelcha@gmail.com

Figura 1: Haru: El Robot Social.

Para interactuar con los humanos, Haru está equipado con un sistema de sensores sofisticado. Uno de estos sensores es la cámara RGBD Azure Kinect. Este dispositivo desarrollado por Microsoft, además de los incorporados a bordo de del robot, hace que Haru cuente con un elaborado sistema de percepción que permite reconocer a las personas, seguir sus movimientos y responder de manera apropiada a sus preguntas.

Este estudio se centra en la aplicación de técnicas de aprendizaje automático a datos de esqueletos recogidos mediante la cámara RGBD Azure Kinect con el propósito de ampliar el sistema de percepción del robot. El artículo se organiza de la siguiente manera: en la Sección 2 se describen los fundamentos y trabajos relacionados. En la Sección 3 se describe la metodología para la recopilación y procesamiento de datos, así como de los rasgos fundamentales empleados para la detección de F-formations. La Sección 4 presenta resultados experimentales de la aplicación de diversos modelos de aprendizaje automático. Finalmente, la Sección 5 presenta las conclusiones y trabajos futuros.

2. Fundamentos y trabajos relacionados

2.1. Concepto de F-Formation

Las F-Formations, introducidas por el reconocido etnometodólogo Adam Kendon (Kendon, 1990), son patrones espaciales que surgen cuando un grupo de personas interactúa entre sí. Estas formaciones se caracterizan por la disposición relativa de los individuos, que varía en términos de proximidad y orientación. Además, dentro de las F-Formations, se pueden distinguir diferentes tipos de zonas proxémicas, como la zona íntima (Zona O), la zona personal (Zona P) y la zona ajena a la conversación que se sitúa a la espalda de los participantes (Zona R) que reflejan los niveles de cercanía y familiaridad entre los miembros del grupo (Figura 2).

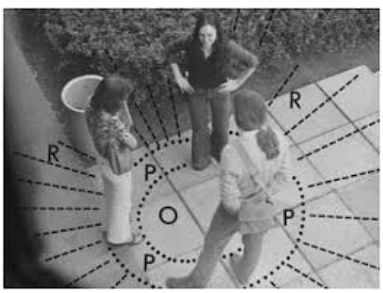

Figura 2: F-Formation entre tres personas. Se indican las zonas O, P y R

La identificación y comprensión de las F-Formations, incluyendo las distintas zonas, es de gran ayuda a la hora de crear sistemas de planificación de trayectorias para robots sociales móviles, o como método de evaluación para saber cuando una máquina podría incorporarse o participar en una conversación que tiene lugar entre humanos.

2.2. Sistema subyacente de fusión de datos

El robot Haru está equipado con un sistema de fusión de datos que combina información de diversos sensores. Este sistema captura datos de sensores como la cámara Azure Kinect, el micrófono y los sensores de movimiento del robot, pero también puede incorporar información de otros sensores especializados, como aquellos dedicados a la detección de manos. Esto permite la realización de estimaciones más precisas basadas en la conjunción de diferentes tipos de datos.

La fusión de estos datos da como resultado un mensaje `People` de ROS (Ragel et al., 2022), que contiene información de cada persona presente en la escena y cuyo contenido se muestra en la Figura 3. Los datos sobre cada persona en la figura corresponden a su expresión facial.

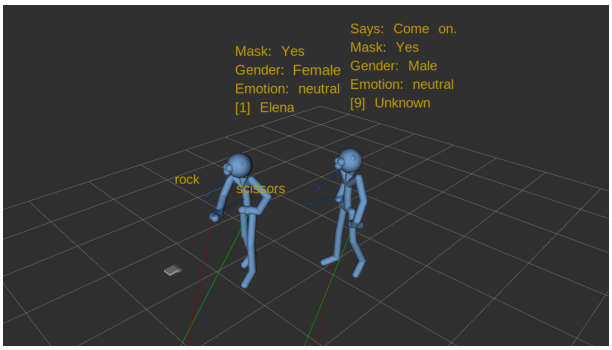

Figura 3: Visualización del mensaje `People`. Éste mensaje incluye información del esqueleto de las personas en el campo de visión.

Este sistema de fusión de datos proporciona a Haru una visión más completa y contextualizada de los grupos conversacionales, al poder detectar también las características individuales de las personas en escena. Además, mejora la adaptabilidad del robot en entornos sociales dinámicos, permitiendo por ejemplo una planificación de trayectorias más eficiente en el caso de robots móviles.

2.3. Otros métodos de detección de F-formations

Se han propuesto diferentes métodos para la detección de F-Formations, la mayoría de ellos basados en el reconocimiento por imagen. Por ejemplo, Hung and Kr.ºse (2011) presentaron un enfoque que utiliza el conjunto dominante para detectar F-Formations en imágenes. Cristani et al. (2011) llevaron a cabo un análisis estadístico de las F-Formations para descubrir interacciones sociales. Setti et al. (2015) propusieron un método de detección de grupos a múltiples escalas. Swofford et al. (2020) introdujeron DANTE, una Deep Affinity Network para agrupar individuos en conversaciones. Por último, Thompson et al. (2021) se basaron en DANTE y exploraron la detección de grupos conversacionales utilizando redes neuronales de grafos.

El presente estudio adopta un enfoque distinto, centrado en el análisis de esqueletos 3D. La manipulación de estos datos, extraídos de un ”bag”, resulta intrínsecamente menos pesada

en comparación con conjuntos de imágenes, lo cual disminuye drásticamente la cantidad de datos requeridos para el entrenamiento de los modelos. Esta optimización de la eficiencia computacional, unida a la robustez que aporta frente a variaciones de iluminación y oclusión, representa una evolución en nuestra metodología.

3. Metodología

3.1. Recopilación de datos

Para llevar a cabo la adquisición de datos, se implementó un procedimiento estructurado compuesto por varias fases.

People[]	
Person	
`Person ID`	
`Skeleton`	`Body Parts[]` `Body Gestures[]`

`Face`	`Name`
	`Gender`
	`Mask`
	`Emotions[]`
	`Bounding Box`

`Right Hand`	`Hand Parts[]`
	`Gestures[]`
	`Direction`

`Left Hand`	`Hand Parts[]`
	`Gestures[]`
	`Direction`

`Speech`	`Transcription`
	`Transcription Confidence`
	`Speaker Confidence`
	`Wake-up Word`

Figura 4: Estructura del mensaje `People` capturado por el sistema de Haru

En la fase inicial, se configuraron escenas representativas de grupos conversacionales de dos y tres participantes. Estas escenas fueron cuidadosamente monitoreadas y grabadas. Cada instancia de grabación fue categorizada en base a la presencia o ausencia de formación de grupos conversacionales, es decir, etiquetada como `True` si es una escena con grupos y `False` si las personas en escena no forman un grupo.

Para el registro de los datos, se utilizó la herramienta rosbag, integrada en el entorno de ROS. Esta herramienta permitió capturar y almacenar la información proveniente del topic `People` (Figura 4), que proporcionaba datos críticos para el estudio.

Cada persona publicada en el topic `people` tiene una gran cantidad de datos asociados, vease **?**, en una primera instancia tenemos que filtrar los datos de interés que en este caso serán los atributos `ID` y `Body Parts[]` de Skeleton.

La lista de `Body Parts []` esta compuesta por treinta y dos partes del cuerpo (Figura 5), cada una tiene una posición tridimensional y una orientación dada por un cuaternión.

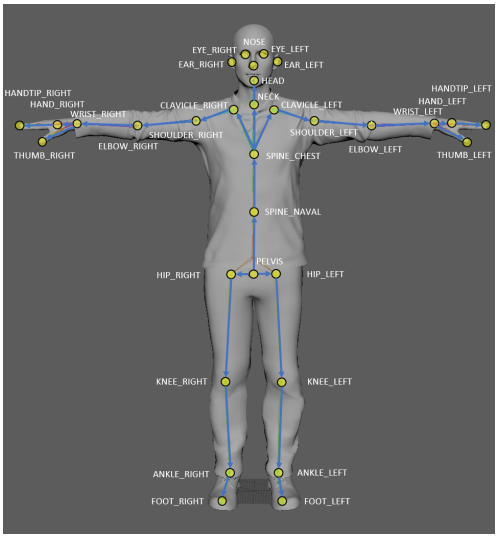

Figura 5: Partes del cuerpo que provee para cada persona el sistema de fusión de datos, enviados en el mensaje `People`.

3.2. Preprocesamiento de datos

En la primera fase de procesamiento, realizamos una transformación de los datos capturados, agrupando por instante de tiempo los esqueletos detectados en la escena. Esto resulta en que cada instancia de los datos para el entrenamiento consta de una vector de los `ID`, junto con matrices que recogen posiciones y orientaciones de cada parte del cuerpo. Ej: Feature de una instancia de entrenamiento como puede ser la posición de la cabeza contendrá las posiciones de las personas presentes en escena.

Sin embargo, características que provienen de posición y de orientación requieren de un preprocesamiento distinto que será explicado en los siguientes dos apartados.

3.2.1. Preprocesamiento de posiciones

En esta operación, se calculan las distancias relativas entre cada parte del cuerpo p de n personas presentes en la escena. Se utiliza la fórmula de distancia euclídea en el plano XY para calcular la distancia entre dos puntos.

Estas distancias se organizan en una matriz triangular superior D, donde cada elemento D_{ij} representa la distancia entre las partes del cuerpo p_i y p_j.

$$D_{n\times n} = \begin{bmatrix} 0 & d(\mathbf{p}_1,\mathbf{p}_2) & d(\mathbf{p}_1,\mathbf{p}_3) & \cdots & d(\mathbf{p}_1,\mathbf{p}_n) \\ & 0 & d(\mathbf{p}_2,\mathbf{p}_3) & \cdots & d(\mathbf{p}_2,\mathbf{p}_n) \\ & & \ddots & \ddots & \vdots \\ & & & 0 & d(\mathbf{p}_{n-1},\mathbf{p}_n) \\ & & & & 0 \end{bmatrix} \quad (1)$$

Finalmente, se aplana la matriz triangular superior D en un vector unidimensional L, donde se concatenan todas las distancias calculadas.

3.2.2. Preprocesamiento de orientaciones

En esta operación, se calculan las orientaciones relativas entre cada parte del cuerpo p de n personas presentes en la escena.

Si cada parte del cuerpo p contiene una orientación dada por un cuaternión q, queremos calcular la diferencia angular relativa del *yaw* para representar las orientaciones relativas en el

eje Z. Se calcula la diferencia entre las rotaciones respecto al eje Z entre dos cuaterniones, y se ajusta al rango de 0 a π. Cada cuaternión $\mathbf{q}_i$ se puede representar como un cuaternión de Hamilton con componentes w_i, x_i, y_i y z_i:

$$\mathbf{q}_i = w_i + x_i i + y_i j + z_i k \tag{2}$$

Para calcular el ángulo respecto al eje Z de un cuaternion, se utiliza la fórmula:

$$\theta = 2 \arctan\left(\frac{y_i}{w_i}\right) \tag{3}$$

La diferencia de ángulos de rotación entre los cuaterniones $\mathbf{q}_i$ y $\mathbf{q}_j$ se calcula como:

$$\Delta\theta_{ij} = |\theta_i - \theta_j| \text{ mód } (2\pi) \tag{4}$$

Para asegurarse de que la diferencia de ángulos esté en el rango de 0 a π, se aplica el siguiente ajuste:

$$\Delta\theta_{ij} = \text{mín}(\Delta\theta_{ij}, 2\pi - \Delta\theta_{ij}) \tag{5}$$

Estas distancias se organizan en una matriz triangular superior O, donde cada elemento O_{ij} representa la orientación relativa entre las partes del cuerpo p_i y p_j.

$$O_{n\times n} = \begin{bmatrix} 0 & \Delta\theta_{12} & \Delta\theta_{13} & \cdots & \Delta\theta_{1n} \\ & 0 & \Delta\theta_{23} & \cdots & \Delta\theta_{2n} \\ & & \ddots & \ddots & \vdots \\ & & & 0 & \Delta\theta_{(n-1)n} \\ & & & & 0 \end{bmatrix} \tag{6}$$

Finalmente, se transforma la matriz triangular superior O en un vector unidimensional L, donde se concatenan todas las diferencias de orientación calculadas.

Tras aplicar estas transformaciones, los datos se encuentran listos como datos de entrada para el entrenamiento de modelos de clasificación.

4. Resultados Experimentales

Se ha diseñado un conjunto de experimentos con el fin de validar la metodología propuesta. Así, se han grabado datos de escenarios en los que personas interaccionan en diversas configuraciones: grupos de dos, grupos de tres, individuales, grupos de dos e individuales, etc. El objetivo es analizar si nuestra propuesta de pre-procesamiento de datos y análisis de múltiples partes del cuerpo de cada individuo mejora las tasas de detección de los diferentes grupos en cada escena respecto a aproximaciones del estado del arte. Compararemos nuestra solución con respecto a las características empleadas en otros trabajos en la literatura (Thompson et al., 2021; Swofford et al., 2020), que estiman grupos conversacionales mediante grafos, donde cada nodo del grafo es una persona con una orientación y una posición, una característica en nuestro caso.

Con el fin de llevar a cabo un análisis comparativo exhaustivo, se han entrenado siete modelos diferentes, todos ellos buenos candidatos para llevar a cabo clasificación sobre secuencias de características: Decision Tree, Gaussian Naive Bayes, Gradient Boosting, Logistic Regression, Multi-Layered Perceptron (MLP), Random Forest y Support Vector Machine Classifier (SVC). Estos modelos se han entrenado utilizando sólo la posición/orientación relativa del torso de una persona respecto al de las demás(característica que usan otras aproximaciones del estado del arte) y utilizando todos las partes de cada individuo (nuestra propuesta).

Los párrafos a continuación describen el conjunto de datos utilizado para entrenamiento y validación, y los resultados obtenidos.

4.1. Descripción de los datos

Se han recogido doce secuencias de datos con el objetivo de realizar dos modelos de clasificación de interacciones entre grupos de dos y tres individuos. Seis de estas secuencias retratan interacciones binarias y los restantes interacciones ternarias. Cada subconjunto se dividió en lotes de entrenamiento (cuatro secuencias) y validación (dos secuencias).

El subconjunto de datos correspondiente a los grupos de dos individuos incluye tres secuencias que registran la interacción natural entre dos personas, así como tres secuencias donde las personas se mantienen aisladas, caminando y mirando en direcciones opuestas.

Respecto al subconjunto de datos de grupos de tres personas, se han capturado tres secuencias que documentan interacciones entre tres individuos con la inclusión de grabaciones panorámicas de su entorno. Las situaciones restantes comprenden tres secuencias en los cuales se representa una dinámica de tres individuos no interactuando directamente, dos personas en conversación con una tercera caminando alrededor, y un escenario de rotación de diálogo por parejas entre las tres personas.

Se implementaron las operaciones de preprocesamiento de datos descritas anteriormente sobre cada una de las instancias presentes en los secuencias. Así, para cada instante de tiempo, se obtuvo un vector unidimensional de características junto con su etiqueta (forman o no forman grupo). Este vector se compone de la concatenación de las matrices triangulares superiores aplanadas, correspondientes a las matrices de distancias y orientaciones relativas de cada parte del cuerpo de los individuos presentes. Con ello, se ha creado un conjunto de datos robusto y representativo para el entrenamiento y validación de los modelos.

4.2. Resultados

La Tabla 1 muestra los valores de Precision y Recall para cada uno de los clasificadores sobre el conjunto de validación. Se han llevado a cabo cuatro entrenamientos por cada uno de los siete clasificadores: Grupo de dos con una característica por persona, Grupo de dos con múltiples características por persona, Grupo de tres con una característica por persona y Grupo de tres con múltiples características por persona.

Los valores de Precision y Recall de las aproximaciones utilizando múltiples características son superiores a las de característica única, tal como lo indican los resultados presentados en la tabla 1.

Se puede apreciar en la evolución de la curva ROC (Característica Operativa del Receptor) de cada grupo de clasificadores para grupos de dos personas (Fig. 6) y grupos de tres personas (Fig. 7), que nuestra aproximación ofrece curvas ROC definidas y agudas de 0,0 a 0,2 en el eje X, especialmente para Random Forest y SVC, según lo evidenciado en las Figuras 6 y 7.

	Grupo de dos				Grupo de tres			
	Una caract.		Multi caract.		Una caract.		Multi caract.	
Clasificador	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**
LogisticRegression	0.767	0.385	**0.982**	**0.982**	0.918	0.906	0.924	0.918
DecisionTreeClass	0.320	0.305	0.889	0.888	0.497	0.546	**0.993**	0.993
RandomForestClass	0.340	0.325	0.970	0.968	0.674	0.682	0.989	0.989
SVC	0.130	0.171	0.968	0.965	0.823	0.754	0.717	0.624
MLPClass	0.141	0.201	0.970	0.969	0.711	0.711	0.856	0.818
GaussianNB	0.321	0.283	0.898	0.868	0.946	0.943	0.848	0.847
GradientBoostingClass	0.195	0.260	0.952	0.946	0.760	0.761	0.993	**0.993**

Tabla 1: Precisión y Recall de los clasificadores entrenados.

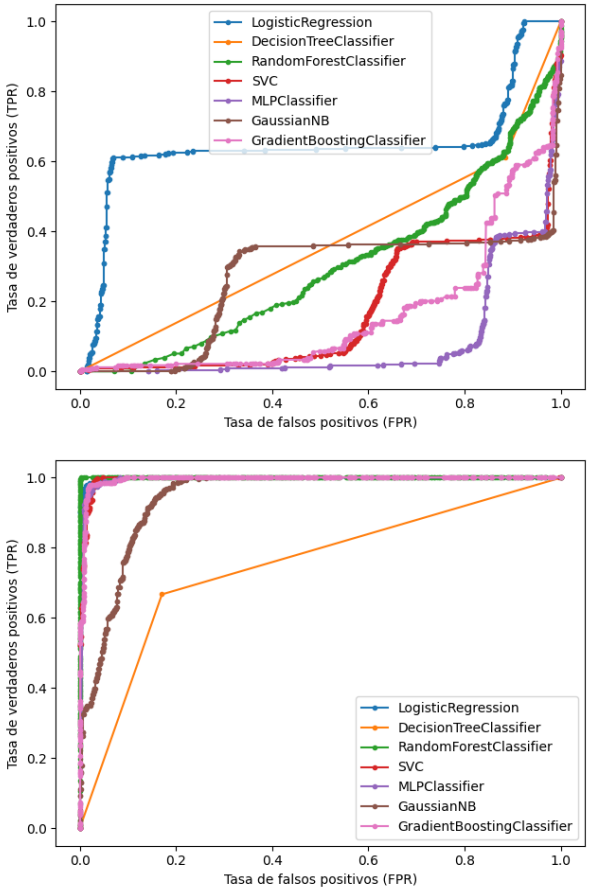

Figura 6: Curvas ROC de los clasificadores de grupos de dos personas. (Arriba) Clasificadores utilizando una sola característica. (Abajo) Clasificadores utilizando múltiples características.

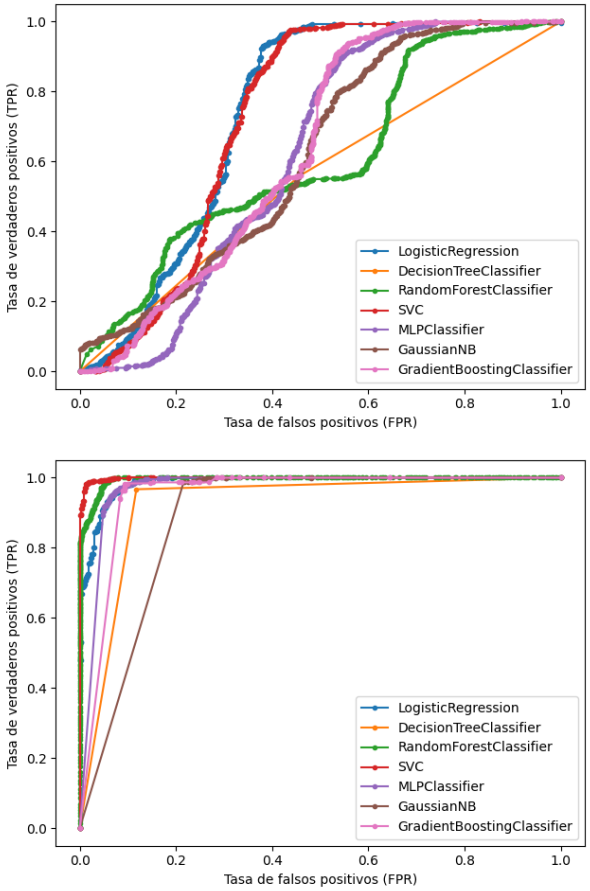

Figura 7: Curvas ROC de los clasificadores de grupos de tres personas. (Arriba) Clasificadores utilizando una sola característica. (Abajo) Clasificadores utilizando múltiples características.

Estos resultados concuerdan con la lógica de que un número mayor de características permiten una mayor redundancia frente a oclusiones, ruido o incluso errores en el detector de personas, facilitando la detección de grupos de personas. Más aún, la posición y orientación del torso (solo una característica) puede llegar a ser confusa a la hora de determinar grupos, dado que las personas pueden estar agrupadas en orientaciones no necesariamente alineadas, dado que la conversación se fundamenta en el contacto visual.

Durante las pruebas realizadas en entornos reales, se corroboró que los clasificadores con rendimientos más óptimos para identificar grupos de dos personas y grupos de tres personas fueron MLP y RandomForest, respectivamente. La Figura 8 ilustra las salidas de estos clasificadores.

Para profundizar en el análisis de estos modelos, aplicamos un test de permutaciones al clasificador MLP y extrajimos la importancia de las características en el RandomForest. Ambos estudios revelaron que estos modelos asignan un peso o importancia significativa a las partes del cuerpo relacionadas con el busto, como la cabeza, las clavículas y el cuello.

En particular, las orientaciones de los ojos y la nariz fueron identificadas como las características más relevantes a la hora

de determinar si un grupo de personas está o no en una conversación. Los resultados sugieren que los indicadores visuales pueden jugar un papel crucial en la detección de grupos conversacionales y subrayan la capacidad de nuestros clasificadores para interpretar estas señales en entornos prácticos.

En resumen, los resultados obtenidos indican que la combinación de múltiples características corporales y su análisis por los clasificadores MLP y RandomForest, proporciona una identificación efectiva y robusta de grupos de personas en conversación en entornos reales.

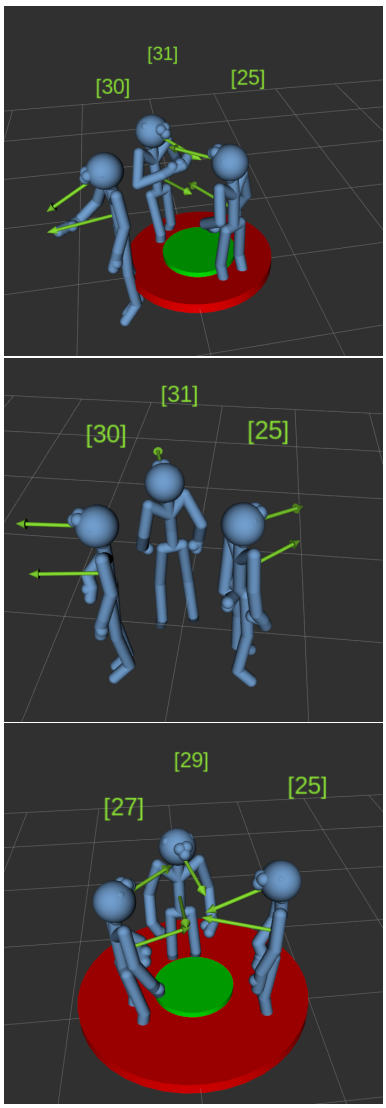

Figura 8: Salida del sistema. Grupo de dos personas siendo detectado (arriba). No se detectan grupos (centro). Grupo de tres personas siendo detectado(abajo)

5. Conclusiones

Este estudio ha permitido entender mejor el impacto significativo que tienen las características corporales en la detección de grupos de personas en conversación. Los resultados indican que la combinación de múltiples características, especialmente aquellas relacionadas con el busto y la orientación de los ojos y la nariz, pueden mejorar la precisión de los clasificadores. Por tanto, un punto de estudio importante dentro de las F-Formations podría ser el análisis del lenguaje corporal de las personas en el entorno, con un enfoque más individual.

En términos de mejoras futuras, se podría explorar la combinación de la tecnología propuesta con métodos y tecnología que puedan ofrecer una visión más amplia de un conjunto de personas, como los métodos basados en grafos o el uso de modelos de inteligencia artificial más sofisticados, que puedan ayudar a sobrevenir los posibles inconvenientes derivados del limitado rango de visión que normalmente poseen las cámaras con las que se plantean este tipo de objetivos.

Agradecimientos

Este trabajo está parcialmente soportado por el Programa Operativo FEDER Andalucía 2014-2020, Consejería de Economía, Conocimiento y Universidades (DeepBot, PY20_00817) y el proyecto NHoA PLEC2021-007868, financiado por el MCIN/AEI/10.13039/501100011033 y la Unión Europea NextGenerationEU/PRTR.

Referencias

Cristani, M., Bazzani, L., Paggetti, G., Fossati, A., Tosato, D., Del Bue, A., Menegaz, G., Murino, V., 2011. Social interaction discovery by statistical analysis of f-formations. In: Proceedings of the British Machine Vision Conference.

Gomez, R., Nakamura, K., Szapiro, D., Merino, L., 2020. A Holistic Approach in Designing Tabletop Robot's Expressivity. pp. 1970–1976. DOI: 10.1109/ICRA40945.2020.9197016

Hung, H., Kr.ºse, B., 2011. Detecting f-formations as dominant sets. In: Proceedings of the 13th international conference on multimodal interfaces. pp. 231–238.

Kendon, A., 1990. Conducting interaction: Patterns of behavior in focused encounters. Cambridge University Press.

Ragel, R., et al., 2022. Multi-modal data fusion for people perception in the social robot haru. In: Social Robotics. ICSR 2022. Lecture Notes in Computer Science. Vol. 13817. Springer, Cham. DOI: 10.1007/978-3-031-24667-8$_1$6

Setti, F., Lanz, O., Ferrario, R., Murino, V., Cristani, M., 2015. Multi-scale f-formation discovery for group detection. In: Proceedings of the IEEE International Conference on Image Processing.

Swofford, M., Peruzzi, J., Tsoi, N., Thompson, S., Mart'in-Mart'in, R., Savarese, S., V'azquez, M., 2020. Improving social awareness through dante: A deep affinity network for clustering conversational interactants. Journal of the Association for Computing Machinery 37 (4). DOI: 10.1145/3412781

Thompson, S., Gupta, A., Gupta, A. W., Chen, A., V'azquez, M., 2021. Conversational group detection with graph neural networks. In: ICMI '21: Proceedings of the 2021 International Conference on Multimodal Interaction. pp. 1–9.