Universidad Nacional Mayor de San Marcos
Universidad del Perú. Decana de América
Facultad de Ingeniería de Sistemas e Informática
Escuela Profesional de Ingeniería de Sistemas

Reconocimiento de gestos dinámicos de brazos en tiempo real para la implementación de un traductor de lengua de señas mediante cámaras de profundidad

TESIS

Para optar el Título Profesional de Ingeniero de Sistemas

AUTOR

Pedro Emilio VARGAS PABLO

ASESOR

Mg. Augusto Parcemon CORTEZ VÁSQUEZ

Lima, Perú

2019

Referencia bibliográfica

Vargas, P. (2019). *Reconocimiento de gestos dinámicos de brazos en tiempo real para la implementación de un traductor de lengua de señas mediante cámaras de profundidad*. [Tesis de pregrado, Universidad Nacional Mayor de San Marcos, Facultad de Ingeniería de Sistemas e Informática, Escuela Profesional de Ingeniería de Sistemas]. Repositorio institucional Cybertesis UNMSM.

Metadatos complementarios autor/ asesor

Datos de autor	
Nombres y apellidos	Pedro Emilio Vargas Pablo
Tipo de documento de identidad	DNI
Número de documento de identidad	73248916
URL de ORCID	
Datos de asesor	
Nombres y apellidos	Augusto Parcemon Cortez Vasquez
Tipo de documento de identidad	DNI
Número de documento de identidad	08634618
URL de ORCID	https://orcid.org/0000-0002-5188-7962
Datos del jurado	
Presidente del jurado	
Nombres y apellidos	Robert Espinoza Dominguez
Tipo de documento	DNI
Número de documento de identidad	08136325
Miembro del jurado 1	
Nombres y apellidos	Luzmila Pró Concepción
Tipo de documento	DNI
Número de documento de identidad	08862360
Datos de investigación	
Línea de investigación	C.0.3.16 Procesamiento de Imágenes y Visión por computadora
Grupo de investigación	No aplica
Agencia de financiamiento	Sin financiamiento
Ubicación geográfica de la investigación	País: Perú Departamento: Lima Provincia: Huarochirí Distrito: Santo Domingo De Los Olleros Centro poblado: Pucará Calle: Av. Santo Domingo Latitud: -12.222667 Longitud: -76.843056
Año o rango de años en que se realizó la investigación	2016 - 2018
URL de disciplinas OCDE	Otras ingenierías y tecnologías https://purl.org/pe-repo/ocde/ford#2.11.02

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS
FACULTAD DE INGENIERIA DE SISTEMAS E INFORMATICA
Escuela Académico Profesional de Ingeniería de Sistemas

Acta de Sustentación de Tesis

Siendo las 14:20 horas del día 05 de abril del año 2019 se reunieron los docentes designados como miembros de Jurado de Tesis, presidido por el Ing. Robert Espinoza Domínguez (Presidente), la Dra. Luzmila E. Pró Concepción (Miembro) y Mg. Augusto P. Cortez Vásquez (Miembro Asesor) para la sustentación de la Tesis Intitulada: **"Reconocimiento de gestos dinámicos de brazos en tiempo real para la implementación de un traductor de lengua de señas mediante cámaras de profundidad".** Del Bachiller: **Pedro Emilio Vargas Pablo;** para obtener el Título Profesional de Ingeniero de Sistemas.

Acto seguido de la exposición de la Tesis, el Presidente invitó al Bachiller a dar las respuestas a las preguntas establecidas por los Miembros del Jurado.

El Bachiller en el curso de sus intervenciones demostró pleno dominio del tema, al responder con acierto y fluidez a las observaciones y preguntas formuladas por los señores miembros del Jurado.

Finalmente habiéndose efectuado la calificación correspondiente por los miembros del Jurado, el bachiller obtuvo la nota de 19 (En letras) DIECINUEVE

A continuación el Presidente del Jurado Ing. Robert Espinoza Domínguez, declara al Bachiller **Ingeniero de Sistemas.**

Siendo las 15:20 horas, se levantó la sesión.

.....
Presidente
Ing. Robert Espinoza Domínguez

.....
Miembro
Dra. Luzmila E. Pró Concepción

.....
Miembro Asesor
Mg. Augusto P. Cortez Vásquez

FICHA CATALOGRÁFICA

TÍTULO DE LA TESIS:

Reconocimiento de Gestos Dinámicos de Brazos en Tiempo Real para la Implementación de un Traductor de Lengua de Señas mediante Cámaras de Profundidad

AUTOR: Vargas Pablo, Pedro Emilio

ASESOR: Cortez Vásquez, Augusto

ÁREA / PROGRAMA / LÍNEA DE INVESTIGACIÓN: Ingenierías / Diseño y Aplicación de Nuevas Tecnologías / Inteligencia Artificial

(Lima, Perú 2019)

Tesis, Facultad de Ingeniería de Sistemas, Pregrado, Universidad Nacional Mayor De San Marcos

Formato: 28 x 20 cm Páginas: xix, 174

2019

DEDICATORIA

Este trabajo está dedicado a quienes luchan por sus sueños y no se amilanan ante la rudeza de los obstáculos que se ven obligados a superar y de los cuales aprenden en pro de ser mejores en sus diferentes perfiles.

Y, no con menor importancia, a quienes han depositado su confianza en mi persona.

AGRADECIMIENTOS

Agradezco infinitamente a mis padres por su confianza, quienes hicieron sacrificios en diferentes oportunidades con tal de verme lograr mis sueños, renunciando a sus necesidades en los momentos más difíciles que nos ha tocado vivir y superar como familia. En este ámbito, hago especial mención a la maestra Jenny Vega Lázaro, actual pareja de mi padre a quien debo la oportunidad de prepararme, ingresar y llevar mi carrera profesional en esta gloriosa universidad, persona la cual me abrió las puertas de su hogar y familia, brindándome no sólo el cariño y aceptación de ellos, sino también de sus hijos, mis hermanos.

A mis parientes, quienes forjaron en mí una personalidad con la convicción y fortaleza necesaria para no doblegarme en momentos cruciales, permitiendo desarrollarme como una persona y profesional con valores y capacidades arraigadas en beneficio y orgullo de mi nación.

Al Magister Augusto Cortez Vásquez por su tiempo, orientación y dedicación para que este trabajo cumpla con los objetivos trazados, así como los consejos brindados para mejorar como persona y profesional en formación.

A Otto Ángel Castillejo Melgarejo, de quien recibí apoyo desinteresado para permitirme conocer la cultura de los no oyentes, sus principales preocupaciones, problemas y metas sociales.

A los colegas y amigos hechos a lo largo de los años de estudio, quienes se convirtieron en un incentivo más para progresar día tras día.

A todas aquellas personas que indirectamente me ayudaron a culminar este trabajo y quienes muchas veces constituyeron un invalorable apoyo.

Y por encima de todo doy gracias a Dios por enrumbarme en el camino correcto, siendo el guía y aliento que mi alma necesitó en varias oportunidades.

UNIVERSIDAD NACIONAL MAYOR DE SAN MARCOS

UNIVERSIDAD DEL PERÚ, DECANA DE AMERICA

FACULTAD DE INGENIERIA DE SISTEMAS E INFORMATICA

ESCUELA PROFESIONAL DE INGENIERIA DE SISTEMAS

RECONOCIMIENTO DE GESTOS DINÁMICOS DE BRAZOS EN TIEMPO REAL PARA LA IMPLEMENTACIÓN DE UN TRADUCTOR DE LENGUA DE SEÑAS MEDIANTE CÁMARAS DE PROFUNDIDAD

AUTOR : VARGAS PABLO, PEDRO EMILIO

ASESOR : CORTEZ VÁSQUEZ, AUGUSTO

TÍTULO : TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS

FECHA : ABRIL 2019

RESUMEN

Según la World Federation of the Deaf, existen aproximadamente 70 millones de personas a nivel mundial con deficiencias auditivas, de ellas un 80% no tiene acceso a la educación y sólo 1 a 2% cuenta con formación en Lengua de Señas como medio de comunicación. Sin embargo, enfrentan obstáculos para su desarrollo en la sociedad, por lo cual se han establecido normativas a nivel mundial, pero en la práctica no son acatadas por las entidades a pesar de su obligatoriedad. Una solución propuesta por el gobierno nacional es ofrecer servicios de intérpretes como mediadores y facilitadores, sin embargo, para el año 2013 sólo habían sido capaces de atender a no más del 10% de solicitantes, considerando además que el servicio cuenta con un horario restringido y un trámite lento. Frente a ello, allanar obstáculos de comunicación mediante un software traductor sería un gran aporte social, supliendo en cierta medida el rol de los intérpretes y abriendo puertas a quienes deseen superarse. Un tipo de planteamiento con notable actividad en los últimos años es el reconocimiento de gestos mediante software basándose principalmente en la obtención y procesamiento de datos a partir de imágenes de cámaras RGB y el empleo de métodos probabilísticos (Principalmente HMM y Redes Neuronales), generando altos costos computacionales y requiriendo mayor tiempo de desarrollo a cambio de una tasa de reconocimiento aceptable. Como consecuencia, esta tesis propone el empleo de data 3D a partir de una cámara de profundidad, empleando DTW como método clasificador para el reconocimiento de gestos. El presente proyecto ha logrado un porcentaje de reconocimiento del 98.18%.

Palabras Claves: Reconocimiento de Gestos, Kinect, DTW, Cámara de profundidad.

SAN MARCOS NATIONAL MAJOR UNIVERSITY
UNIVERSITY OF PERU, DEAN OF AMERICA
SYSTEM ENGINEERING AND INFORMATIC FACULTY
SYSTEM ENGINEERING PROFESSIONAL COLLEGUE

RECOGNITION OF DYNAMIC GESTURES OF ARMS IN REAL TIME FOR THE IMPLEMENTATION OF A SIGN LANGUAGE TRANSLATOR USING DEPTH CAMERAS

AUTHOR : VARGAS PABLO, PEDRO EMILIO

ASSESSOR : CORTEZ VÁSQUEZ, AUGUSTO

GRADE : GRADUATE STUDIES IN SYSTEM ENGINEERING

DATE : APRIL 2019

ABSTRACT

According to the World Federation of the Deaf, approximately 70 million of people around the world have hearing impairments, 80% of them do not have access to education and 1 to 2% learned Sign Language and are able to use it as a way of communication. However, they are facing obstacles for them social development, therefore some normative were passed in all the world, but them are not executed by the entities in spite of the obligatoriness. A solution proposed by the national government is to offer interpreters services as mediators and facilitators; however, at 2013 they had only been able to serve no more than 10% of applicants, also considering that the service has a restricted schedule and a slow process. As consequence, reducing communication obstacles through translator software would be a great social contribution, doing to a certain extent the role of the interpreters and opening chances for those people who wish to overcome oneself. A kind of approach with remarkable activity in recent years is the recognition of gestures using software based mainly on obtaining and processing data from images of RGB cameras and the use of probabilistic methods (Mainly HMM and Neural Networks), generating high computational costs and requiring more development time in exchange for an acceptable recognition rate. For those reasons, this thesis proposes the use of 3D data from a depth camera, using DTW as a classifier method for gesture recognition. This project has achieved a recognition percentage of 98.18%.

Key words: Sign Language Recognition, Kinect, DTW, Depth Camera.

ÍNDICE

LISTA DE FIGURAS

LISTA DE TABLAS

INTRODUCCIÓN

Actualmente existen personas con diversos tipos de discapacidad, condición la cual limita su desarrollo e Inclusión Social. Acorde al tipo de minusvalía, innumerables entes gubernamentales han dictaminado soluciones normativas para sobrellevar la situación de la mejor forma. Así, para el caso de los no oyentes, se tienen más de 70 millones a nivel mundial y poco más de medio millón reside en nuestro país. En este aspecto, las propuestas han consistido en la contratación de intérpretes como forma de reducir brechas en la Inclusión Social, sin embargo, son soluciones con muy limitada capacidad de atención, de largo plazo, elevado costo y riesgosas dada la posibilidad de no ser acatadas.

Tenaces investigadores han sugerido y llevado a cabo la implementación de softwares traductores de lengua de señas como apoyo al rol de intérpretes. Sin embargo, dichas propuestas se ven limitadas en alcance y aún no han sido aplicadas a la Lengua de Señas Peruana. Es así que, después de un análisis sobre la literatura mundial relacionada, en este trabajo se desarrolló un traductor de gestos de brazos y manos en tiempo real, no intrusiva, a reducido costo y una tasa de reconocimiento del 98.18% para gestos definidos en contextos laborales con intenciones de propiciar la Inclusión Social, superando el enfoque y resultados de terceros, pero actuando como paso inicial para lograr en etapas futuras la implementación de un Intérprete de Lengua de Señas.

La tesis está organizada en 6 capítulos, detallados brevemente a continuación:

El Capítulo I declarado como Planteamiento del Problema, hace mención a los antecedentes del problema tratado, así como la definición y justificación del mismo, además se detallan los objetivos y alcance del presente trabajo.

El Capítulo II aborda el Marco Teórico con la finalidad de dotar al lector de conocimientos y conceptos necesarios para comprender sin mayor problema los puntos tratados durante el desarrollo de esta propuesta, así como los términos técnicos empleados.

El Capítulo III comprende el Estado del Arte. Realiza una revisión de los trabajos y propuestas previas encontradas, revisadas y catalogadas como más resaltantes y las cuales resultarán de apoyo en este trabajo, así también una comparación entre los mismos con la

finalidad de ir seleccionando los que se consideren aprovechables para lograr el objetivo planteado.

Mientras, el Capítulo IV titulado Desarrollo de la Solución Propuesta, muestra el desarrollo considerando los componentes software y explicando los pasos desde la captura de datos hasta el output deseado en este proyecto.

Los resultados obtenidos en base a experimentaciones con la propuesta son mostrados en el Capítulo V, además son comparadas con otras investigaciones consideradas en el Estado del Arte.

Finalmente, en el Capítulo VI se plantean las conclusiones obtenidas y se deja especificado el trabajo e implementación que se propone a futuro como continuación del presente proyecto.

CAPÍTULO I: PLANTEAMIENTO DEL PROBLEMA

1.1 Antecedentes

El hombre de las cavernas fue uno de los primeros en emplear gestos para expresarse, tras siglos, la necesidad de comunicación permitió el desarrollo de la voz y posteriormente lo que ahora se conoce como El Habla.

Históricamente han existido personas quienes por limitaciones biológicas se han visto impedidos en su normal desempeño. Dicha situación en el contexto actual se conoce como discapacidad[1], éstas pueden variar desde el ámbito mental hasta aspectos físicos. Así se tienen discapacidades relacionadas con el habla y/o escucha como consecuencia de no desarrollarse o haberlos perdido en alguna etapa de vida [Anexo 01]. Gracias a lo anterior, desde tiempos de Aristóteles se han subestimado las capacidades cognitivas y sociales de estas personas, siendo hasta fines del siglo XVI cuando entes como Girolamno Cardano fueron capaces de demostrar y detractar los estereotipos mencionados. [1]

Pero, dichos prejuicios no se han menguado en la sociedad a pesar de la existencia de normas, instituciones y grupos de personas creados con la finalidad de intervenir y erradicar este tipo de problema [1]. Teniendo en cuenta la problemática, la ONU[2] en la "Convención sobre los Derechos de las Personas con Discapacidad" [2] llevada a cabo el 06 de Diciembre del 2006, reconoce una serie de preámbulos o antecedentes [Nota 01] sociales que los exhorta a establecer y declarar un conjunto de artículos con las intenciones de solucionarlos. Entre ellos se citan dos importantes:

---

[1] Se comprende como la ausencia o limitación para realizar actividades dependientes de facultades mentales o físicas del individuo, ello como consecuencia imposibilita el desarrollo eficiente de la persona sobre sus actividades.

[2] "Organización de las Naciones Unidas" o "Naciones Unidas" (NN. UU.), es la mayor organización internacional existente. Esta entidad tiene por finalidad facilitar la cooperación en asuntos como el Derecho Internacional, la Paz y Seguridad Internacional, el desarrollo económico y social, los asuntos humanitarios y los derechos humanos.

- ✓ “Artículo 1: Promover, proteger y asegurar el goce pleno y en condiciones de igualdad de todos los derechos humanos y libertades fundamentales por todas las personas con discapacidad, y promover el respeto de su dignidad inherente…”
- ✓ “Artículo 9: Impulsar a los Estados Partes para que: Ofrezcan formas de asistencia humana o animal e intermediarios, incluidos guías, lectores e intérpretes profesionales de la lengua de señas, para facilitar el acceso a edificios y otras instalaciones abiertas al público…”

Según la WFD[3], existen aproximadamente 70 millones de personas a nivel mundial con deficiencias auditivas. Desafortunadamente los estereotipos, prejuicios y barreras sociales sumados a la falta de reconocimiento y aceptación de la Lengua de Señas como parte de su cultura, identidad y como principal medio de comunicación, les impiden disfrutar a plenitud los derechos humanos. [3] [4]

La Lengua de Señas es universal y en cada país es el lenguaje natural de las personas sordas. La Lengua de Señas varía y se va moldeando según diversos factores (Cultura, sociedad, historia, religión, entre otros) de país en país de forma similar al lenguaje hablado, por lo cual en cada país existe una Lengua de Señas acorde a la necesidad y contexto de sus usuarios, en consecuencia, existen: Lengua de Señas Americana, Lengua de Señas de Nicaragua, Lengua de Señas Peruana, entre otros. [3] [1]

Históricamente, el Gobierno del Perú ha dado muy poca importancia a las personas sordas y su rol en la sociedad. En el año 1987 se publicó por primera vez de forma oficial un Manual de Lengua de Señas[4], un vocabulario básico y un video. En 1996 se reimprime el material y se entregó otro video que demostraba ser en cierto grado una adaptación de la Lengua de Señas Americana.

El manual fue gratamente recibido, no obstante, durante los años 1998 – 2000 surgen cuestionamientos al no reflejar las señas verdaderamente empleadas por la comunidad de sordos en el Perú.

---

[3] “World Federation of the Deaf”, en español “Federación Mundial de Sordos”.

[4] Material brindado con la finalidad de contribuir con el proceso de apropiación de la Lengua de Señas por parte de la comunidad interesada.

Por su parte, el ahora Ministerio de la Mujer y Poblaciones Vulnerables (MIMP) fundó el 06 de enero de 1999 el CONADIS (Consejo Nacional para la Integración de la Persona con Discapacidad), el cual es un organismo público ejecutor que desarrolla políticas, propone normas y acciones para lograr la integración social, económica y cultural de las personas con discapacidad en general.

Además, en el año 2000 se creó una mesa de trabajo denominada "Manos que Hablan – Proyección Social de la Lengua de Señas Peruana" integrada por personas naturales y jurídicas como la Asociación de Sordos del Perú[5] cuyo fin era establecer la base sobre la que se formaría la Lengua de Señas Peruana, para lo cual determinaron seguir los pasos de otros países en cuanto al reconocimiento de la Lengua de Señas y previamente realizar un estudio sociolingüístico de la Lengua de Señas en el Perú. [5]

En los años 2003 – 2004 la Comisión Especial de Estudios sobre Discapacidad, impulsada por el congresista Javier Diez Canseco, desarrolló un proyecto de ley que buscaba el reconocimiento de la Lengua de Señas Peruana. Sin embargo, no tuvieron la aprobación del congreso quienes argumentaron no contar con suficiente información acerca de la Lengua de Señas empleada a nivel nacional, además que esta carecía de sintaxis y gramática propia, y se basaba sólo en códigos manuales. [5]

Es la preocupación del congresista Michael Urtecho quien en el 2009 acogió la propuesta de la Fundación Personas Sordas del Perú[6] y empezó a trabajar en el proyecto legislativo, logrando el reconocimiento oficial de la Lengua de Señas Peruana. También, en ese año se crea la Asociación de Intérpretes y Guías de Perú (ASISEP) cuyo fin era reunir a las personas orientadas al servicio de interpretación para que reciban capacitación y profesionalización. [5]

En el 2010, dado el contexto y notable necesidad, el Ministerio de Educación emite públicamente de forma oficial una Guía para el aprendizaje de Lengua de Señas Peruana acompañado de un vocabulario básico. [6]

---

[5] Cuyas siglas son A.S.P., asociación sin fines de lucro, fundada inicialmente el 04 de mayo de 1958.

[6] Es una institución sin fines de lucro, comprometida a realizar acciones en busca de mejorar la calidad de vida de las personas sordas, nacionales y los oyentes, a través de proyectos y programas en áreas de salud, educación, trabajo, deporte, cultura, entre otros.

Según el INEI (Instituto Nacional de Estadística e Informática) en su Primera Encuesta Nacional Especializada sobre Discapacidad realizada en el 2012, en el Perú existían 1'575,402 personas con alguna discapacidad, de estas 532,209 presentaban dificultades para oír, aun usando audífonos. Además, se conoció que cerca de un 55.7 por ciento de ellos debe recurrir a alguna forma especial para comunicarse (Gestos, lectura de labios, lenguaje de señas, lápiz y papel, entre otros). [7]

Frente a esta realidad, en el año 2010 y 2012 el Congreso de la República del Perú aprobó los proyectos de ley 29535 (Ley que otorga Reconocimiento a la Lengua de Señas Peruanas) y 29973 (Ley General de la Persona con Discapacidad) respectivamente, reglamentando condiciones de igualdad y derechos como intentos de promover y proteger a las personas con discapacidades en general e incentivando su desarrollo e inclusión plena y efectiva en la vida política, económica, social, cultural y tecnológica; y como reconocimiento oficial de la Lengua de Señas Peruana para ciudadanos peruanos con discapacidad auditiva.

La Ley N° 29535 especifica lo siguiente: "Las entidades e instituciones públicas o privadas que brinden servicios públicos o de atención al público, deben proveer a las personas con discapacidad auditiva, de manera gratuita, en forma progresiva y según lo establezca el reglamento, el servicio de intérprete para sordos cuando éstos lo requieran. Dichas entidades e instituciones permiten que estas personas comparezcan ante ellas con interpretes reconocidos oficialmente".

Los intérpretes[7] son elementales para la comunidad sorda, apoyan en la realización de actividades socio-educativas que van desde conferencias, capacitaciones, charlas y talleres [8], por lo que también es necesario que institutos y universidades impulsen esta carrera con el fin de formarlos correctamente para la sociedad [Nota 02].

No obstante, en el Perú no existe una carrera profesional como tal para intérpretes, la mayoría de personas ha aprendido en la práctica o por necesidad al tener algún familiar sordo, y mediante cursos ofrecidos por ciertas instituciones.

[7] Los Intérpretes para Sordos son aquellas personas con amplios conocimientos de la Lengua de Señas del territorio en el cual se desempeñan, pueden realizar interpretación simultánea del lenguaje hablado a la lengua de señas o viceversa, en especial en actividades oficiales.

Actualmente, la tecnología juega un rol fundamental en nuestro día a día, especialmente en áreas de salud y alimentación, sirviendo de invaluable apoyo a las personas con discapacidad mediante diversos estudios, avances y proyectos. En ese sentido, en los últimos años la Interacción Humano Computadora ha cobrado mayor relevancia debido a su naturaleza multidisciplinaria, abarcando principalmente áreas como Ingeniería, Psicología y Ergonomía, enfocándose no sólo en la funcionalidad, sino también en la usabilidad de interfaces para permitir al usuario realizar sus tareas de forma eficiente y eficaz intuitivamente sin requerir mayor entrenamiento.

La Interacción Humano Computadora (IHC) abarca diferentes áreas de investigación donde resaltan el Reconocimiento de Voz y el Reconocimiento de Lengua de Señas. El Reconocimiento de Lengua de Señas se ha convertido en un área de investigación atractiva para investigadores y desarrolladores en general, especialmente desde que sensores de profundidad como la Kinect fueron puestos en venta a precios accesibles. Sin embargo, la investigación y resultados obtenidos en este rubro aún no alcanzan el nivel de madurez deseado.

Figura 1.1 Línea de tiempo con antecedentes del problema. Fuente: Elaboración propia.

1.2 Definición de la Problemática

A pesar de la normatividad y creación de instituciones de formación técnica, para el 2013 sólo se habían reconocido oficialmente dieciocho intérpretes a nivel nacional según informó la Asociación de Intérpretes del Perú, resultando insuficientes y acudiendo a menos del diez por ciento de la comunidad total [9], quienes además no llegan a gozar de beneficios [Nota 03].

Las empresas y entidades en general se muestran reacias al contrato de intérpretes como facilitadores para la interacción y comunicación con personas que emplean Lengua de Señas.

La carencia de una educación especializada en todos los niveles dificulta superar las barreras sociales de quienes viven esta realidad, imposibilitando su desarrollo e inclusión al nivel de valerse por sus propios medios.

Frente a lo anterior, la disponibilidad de entes que actúen como intérpretes es esencial para que los no oyentes puedan superar las barreras de comunicación y lograr una mayor inclusión. Una comunidad a la cual se le ha prometido Inclusión Social[8], pero dicha terminología ha resultado grande para lo alcanzado.

[8] Es la situación que asegura que todos los ciudadanos sin excepción puedan ejercer sus derechos, aprovechar sus habilidades y tomar ventaja de las oportunidades que encuentran en su medio. ("Ministerio de Desarrollo e Inclusión Social" – MIDIS)

Árbol de Problemas

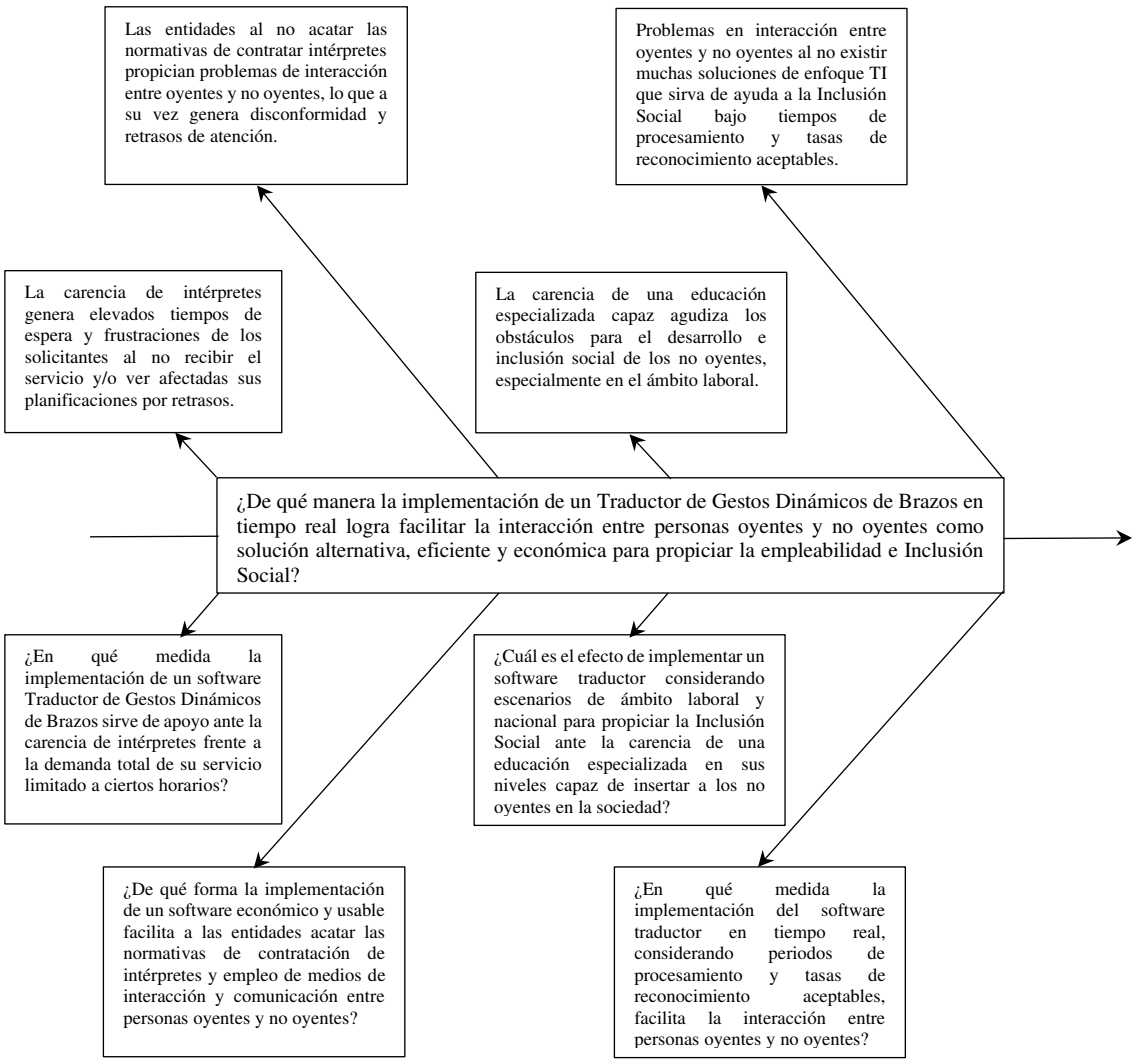


Figura 1.2 Gráfica detallada del árbol de problemas. Fuente: Elaboración propia.

1.3 Formulación del Problema

Problema General

¿De qué manera la implementación de un Traductor de Gestos Dinámicos de Brazos en tiempo real logra facilitar la interacción entre personas oyentes y no oyentes como solución alternativa, eficiente y económica para propiciar la empleabilidad e Inclusión Social?

Problemas Específicos

a) ¿En qué medida la implementación de un software Traductor de Gestos Dinámicos de Brazos sirve de apoyo ante la carencia de intérpretes frente a la demanda total de su servicio limitado a ciertos horarios?

b) ¿De qué forma la implementación de un software económico y usable facilita a las entidades acatar las normativas de contratación de intérpretes y empleo de medios de interacción y comunicación entre personas oyentes y no oyentes?

c) ¿Cuál es el efecto de implementar un software traductor considerando escenarios de ámbito laboral y nacional para propiciar la Inclusión Social ante la carencia de una educación especializada en sus niveles capaz de insertar a los no oyentes en la sociedad?

d) ¿En qué medida la implementación del software traductor en tiempo real, considerando periodos de procesamiento y tasas de reconocimiento aceptables, facilita la interacción entre personas oyentes y no oyentes?

1.4 Formulación de Objetivos

Como consecuencia del contexto actual y problemática indicada, se plantean a continuación los objetivos de la presente investigación.

Objetivo General

Desarrollar un reconocedor de gestos dinámicos de brazos en tiempo real para la implementación de un traductor de lengua de señas, como solución alternativa, eficiente y económica ante la carencia de intérpretes, con el fin de propiciar la empleabilidad e Inclusión Social al facilitar la interacción entre personas oyentes y no oyentes, mediante el uso de cámaras de profundidad.

Objetivos Específicos

a) Evaluar e implementar un software Traductor de Gestos Dinámicos de Brazos que sirva de apoyo ante la carencia de intérpretes frente a la demanda total de su servicio limitado a ciertos horarios.
b) Desarrollar e implementar un software económico y usable que facilite a las entidades acatar las normativas de contratación de intérpretes y empleo de medios de interacción y comunicación entre personas oyentes y no oyentes.
c) Enfocar la implementación del software traductor en escenarios de ámbito laboral y nacional, buscando propiciar la Inclusión Social ante la carencia de una educación especializada en sus niveles capaz de insertar a los no oyentes en la sociedad.
d) Implementar y validar el software traductor en tiempo real considerando periodos de procesamiento y tasas de reconocimiento aceptables para facilitar la interacción entre personas oyentes y no oyentes.

Árbol de Objetivos

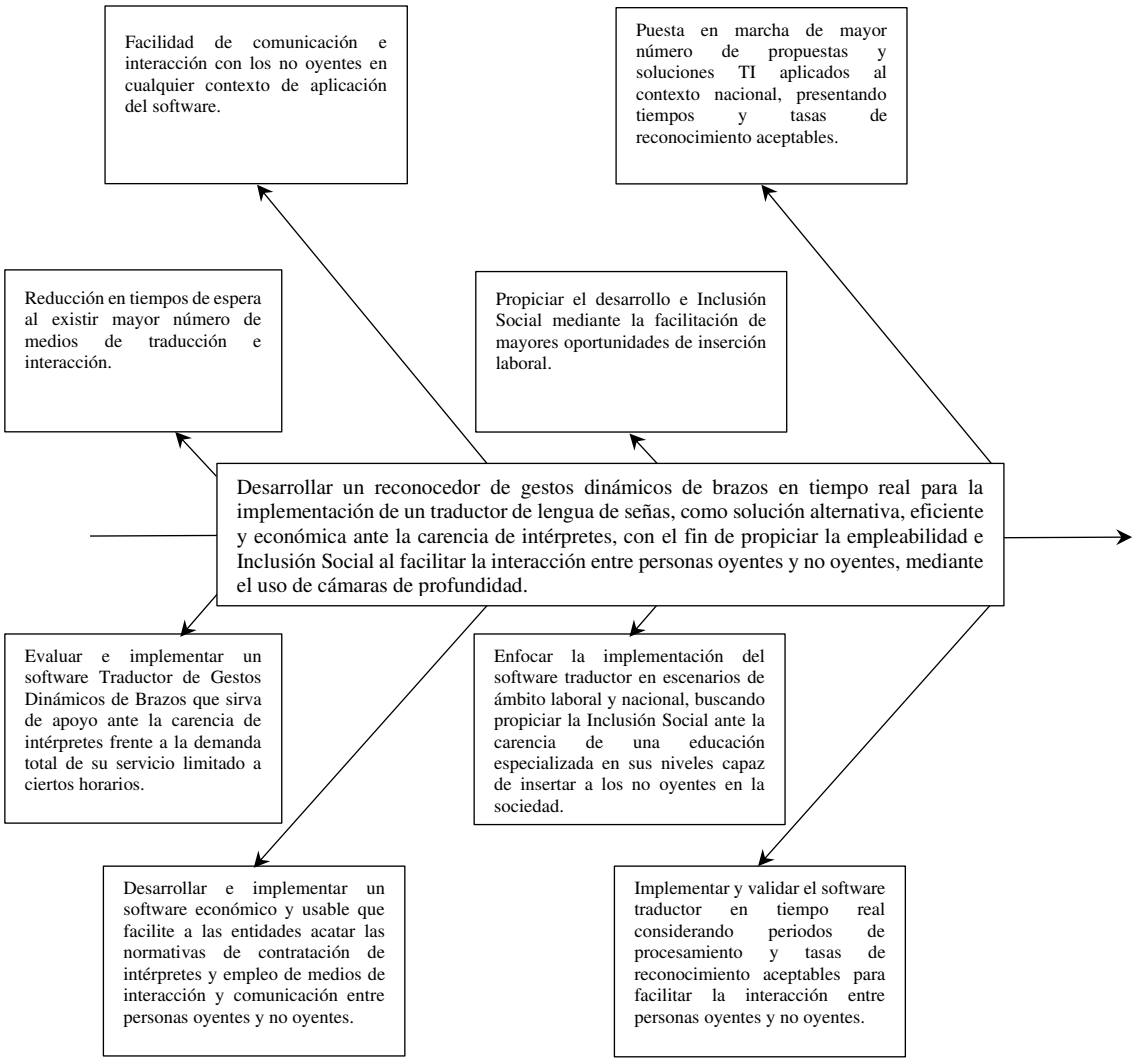


Figura 1.3 Gráfica detallada del árbol de objetivos. Fuente: Elaboración propia.

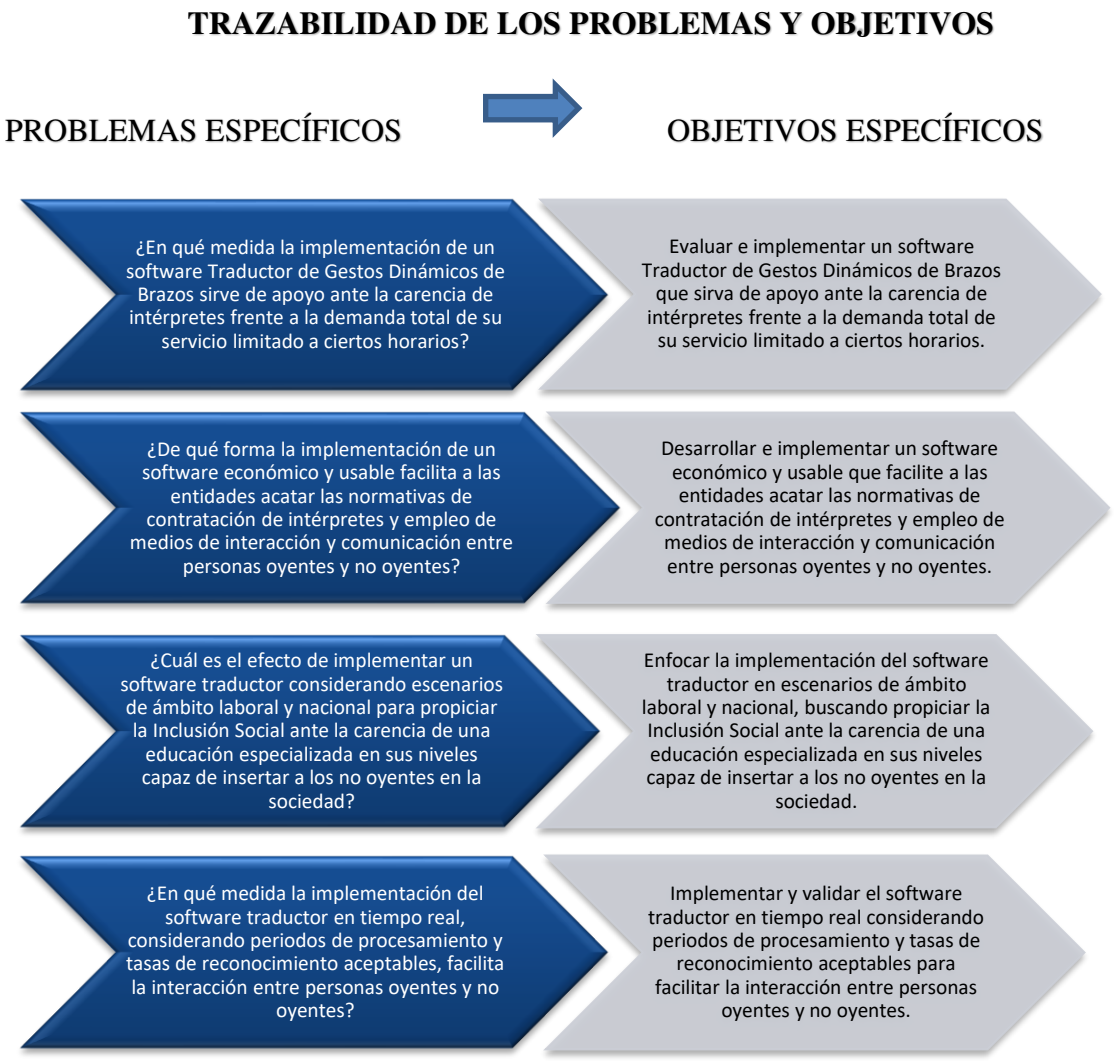


Figura 1.4 Trazabilidad de los problemas y objetivos específicos. Fuente: Elaboración propia.

1.5 Hipótesis y Variables

1.5.1 Hipótesis General

El empleo de un software Traductor de Gestos Dinámicos de Brazos en tiempo real como solución alternativa, eficiente y económica incide favorablemente en la interacción entre personas oyentes y no oyentes ante la carencia de intérpretes, propiciando la empleabilidad e Inclusión Social.

1.6 Identificación de Variables

Variable Independiente

Emplear software Traductor de Gestos Dinámicos de Brazos.

Variable Dependiente

Facilitar mayor interacción entre oyentes y no oyentes lo cual impulsa la empleabilidad e Inclusión Social de los no oyentes.

1.7 Matriz de Consistencia

En este apartado se presenta la Operacionalización de Variables y la Matriz de Consistencia con las cuales se busca mostrar la relación entre el Problema, los Objetivos y la Hipótesis planteada.

1.7.1 Operacionalización de Variables

Para demostrar en qué medida el empleo de un software Traductor de Gestos Dinámicos de Brazos en tiempo real como solución alternativa, eficiente y económica incide favorablemente en la interacción entre personas oyentes y no oyentes, logrando propiciar empleabilidad e Inclusión Social, se procederá a operacionalizar las variables:

VARIABLES	DIMENSIONES	INDICADORES
Variable X: Independiente Emplear software Traductor de Gestos Dinámicos de Brazos.	a) Empleabilidad del software. b) Economía en desarrollo y usabilidad del software. c) Ambientes de uso. d) Eficiencia y eficacia.	Cantidad de gestos ***x1*** Costo económico ***x2*** Usabilidad ***x3*** Contextos de aplicación ***x4*** Tasa de reconocimiento ***x5*** Tiempo de reconocimiento ***x6***
Variable Y: Dependiente Facilitar mayor interacción entre oyentes y no oyentes lo cual impulsa la empleabilidad e Inclusión Social de los no oyentes.	a) Social. b) Laboral. c) Intelectual.	Interacción y comunicación ***y1*** Número de beneficiarios ***y2*** Empleabilidad ***y3*** Accesibilidad y participación en eventos ***y4***

Tabla 1.1 Operacionalización de variables. Fuente: Elaboración propia.

PROBLEMA	OBJETIVOS	HIPÓTESIS	VARIABLES	METODOLOGÍA
Problema General ¿De qué manera la implementación de un Traductor de Gestos Dinámicos de Brazos en tiempo real logra facilitar la interacción entre personas oyentes y no oyentes como solución alternativa, eficiente y económica para propiciar la empleabilidad e Inclusión Social? ***Problemas Específicos*** a) ¿En qué medida la implementación de un software Traductor de Gestos Dinámicos de Brazos sirve de apoyo ante la carencia de intérpretes frente a la demanda total de su servicio limitado a ciertos horarios? b) ¿De qué forma la implementación de un software económico y usable facilita a las entidades acatar las normativas de contratación de intérpretes y empleo de medios de interacción y comunicación entre personas oyentes y no oyentes? c) ¿Cuál es el efecto de implementar un software traductor considerando	***Objetivo General*** Desarrollar un reconocedor de gestos dinámicos de brazos en tiempo real para la implementación de un traductor de lengua de señas, como solución alternativa, eficiente y económica ante la carencia de intérpretes, con el fin de propiciar la empleabilidad e Inclusión Social al facilitar la interacción entre personas oyentes y no oyentes, mediante el uso de cámaras de profundidad. ***Objetivos Específicos*** a) Evaluar e implementar un software Traductor de Gestos Dinámicos de Brazos que sirva de apoyo ante la carencia de intérpretes frente a la demanda total de su servicio limitado a ciertos horarios. b) Desarrollar e implementar un software económico y usable que facilite a las entidades acatar las normativas de contratación de intérpretes y empleo de medios de interacción y comunicación entre personas oyentes y no oyentes. c) Enfocar la implementación del software traductor en escenarios de ámbito laboral y nacional,	***Hipótesis General*** El empleo de un software Traductor de Gestos Dinámicos de Brazos en tiempo real como solución alternativa, eficiente y económica incide favorablemente en la interacción entre personas oyentes y no oyentes ante la carencia de intérpretes, propiciando la empleabilidad e Inclusión Social. ***Hipótesis Específicas*** a) Emplear un software Traductor de Gestos Dinámicos de Brazos servirá de apoyo ante la carencia de intérpretes y la demanda insatisfecha de su servicio limitado a ciertos horarios. b) Implementar y poner en marcha un software económico y usable impulsará a las entidades a acatar las normativas de contratación y empleo de medios de interacción y comunicación entre personas oyentes y no oyentes. c) Desarrollar un software traductor enfocado en el contexto laboral y nacional	***Variables e Indicadores*** Para demostrar y comprobar la hipótesis anteriormente formulada, procederemos a operacionalizarla determinando sus variables e indicadores mostradas a continuación: ***Variable X: Variable Independiente*** Emplear software Traductor de Gestos Dinámicos de Brazos. ***Indicadores:*** Cantidad de gestos ***x1*** Costo económico ***x2*** Usabilidad ***x3*** Contextos de aplicación ***x4*** Tasa de reconocimiento ***x5*** Tiempo de reconocimiento ***x6*** ***Variable Y: Variable Dependiente*** Facilitar mayor interacción entre oyentes y no oyentes lo cual impulsa la empleabilidad e Inclusión Social de los no oyentes. ***Indicadores:*** Interacción y comunicación ***y1***	***Tipo de Investigación*** Por el tipo de investigación mostrado, este trabajo cumple las condiciones metodológicas de una investigación longitudinal. ***Nivel de la Investigación*** Acorde a la naturaleza del estudio de investigación, reúne por su nivel las características de un estudio descriptivo, explicativo y no experimental. ***Método de la Investigación*** Con la finalidad de validar la hipótesis mostrada, se procederán a aplicar los siguientes métodos: **Histórico:** Permite conocer la evolución histórica y hechos que han conllevado al problema de investigación. ***Diseño de la Investigación*** No Experimental

escenarios de ámbito laboral y nacional para propiciar la Inclusión Social ante la carencia de una educación especializada en sus niveles capaz de insertar a los no oyentes en la sociedad? d) ¿En qué medida la implementación del software traductor en tiempo real, considerando periodos de procesamiento y tasas de reconocimiento aceptables, facilita la interacción entre personas oyentes y no oyentes?	buscando propiciar la Inclusión Social ante la carencia de una educación especializada en sus niveles capaz de insertar a los no oyentes en la sociedad. d) Implementar y validar el software traductor en tiempo real considerando periodos de procesamiento y tasas de reconocimiento aceptables para facilitar la interacción entre personas oyentes y no oyentes.	servirá de apoyo para propiciar una mayor Inclusión Social al ser un paso elemental para insertar a los no oyentes en el ámbito laboral. d) Facilitar la interacción entre oyentes y no oyentes, la cual se llevará correctamente mediante la implementación y validación de un software capaz de responder en tiempo real con periodos de procesamiento y tasas de reconocimiento aceptables.	Número de beneficiarios ***y2*** Empleabilidad ***y3*** Accesibilidad y participación en eventos ***y4***	

Tabla 1.2 Matriz de consistencia. Fuente: Elaboración propia.

1.8 Justificación

Según la WFD, aproximadamente el 80% de la población de sordos no tiene acceso a educación y sólo 1 a 2% cuenta con formación en Lengua de Señas. La situación es más crítica en mujeres y niños. Como consecuencia, se considera el desarrollo y reconocimiento legal de Lengua de Señas un requisito fundamental si lo deseado es una participación equitativa en la sociedad de las personas sordas. [3]

En nuestro país la situación es similar, no todas las personas sordas y/o mudas cuentan con las mismas condiciones de educación, tratamiento y rehabilitación. Por ejemplo, dentro de esta comunidad existen quienes pueden leer y escribir, algunos de ellos emplean Lengua de Señas, otros saben leer labios y pocos emplean Lengua de Señas y voz al mismo tiempo para comunicarse. [Anexo 02] [Anexo 03]

Algunos no cuentan con escolaridad, otros sólo primaria completa, unos pocos con secundaria completa y contados los que han alcanzado estudios superiores. [Anexo 04]

En nuestro país estas personas viven aisladas, rechazadas y discriminadas [Anexo 05], viendo frenado su desarrollo. Sólo unos cuantos tienen empleo, además del difícil acceso a la salud y seguridad social. Como resultado, la gran mayoría vive en situación de pobreza, recibiendo marginación y abandono hasta de sus familiares. [Anexo 06]

Si bien se cuenta con la intervención de las autoridades nacionales, es evidente que no alcanzarán grandes resultados a corto plazo o no las lograrán completamente en el largo plazo, no garantizando las condiciones de igualdad anheladas por dicha comunidad.

La comunicación de las personas no oyentes sigue siendo una dificultad con los oyentes (Quienes no hablan Lengua de Señas) al intentar interactuar sin ayuda de algún intérprete del idioma, impidiéndoles realizar sus trabajos o tareas con normalidad.

Una de las recomendaciones propuestas por la WFD es que las sociedades creen sistemas de provisión e igual acceso a los intérpretes de lengua de señas para todas las situaciones en que puedan ser requeridos, sin verse obligados a responder económicamente, sino que dichos servicios deban ser provistos por la sociedad evitando así ser una solución costosa [3]. No obstante, a nivel mundial existen intérpretes profesionales quienes traducen en tiempo real Lengua de Señas, pero los gastos acarreados por viáticos resultan elevados. [10]

Otra solución es contratar intérpretes para que no sólo traduzcan, sino además enseñen a los empleados de empresas e instituciones a comprender lenguas de señas, sin embargo,

es evidente que también es una solución cara y no práctica frente a una propuesta basada en el empleo de tecnología capaz de reconocer Lengua de Señas de forma automática. [4]
En el Perú dicho servicio requiere un trámite documentario, el proceso se restringe a ciertos horarios y días, y además como ya se mencionó, no cuenta con una adecuada proporción de intérpretes frente a las solicitudes.
La presencia de un intérprete que brinde sus servicios al menos gran parte del día facilitaría superar las barreras de comunicaciones y propiciaría el aumento de oportunidades. Pero, dichos intérpretes no siempre están disponibles, por lo que un sistema de reconocimiento automático, barato y con gran disponibilidad es deseable y necesario. [10]
Es por lo mencionado que allanar los obstáculos de comunicación con las personas que emplean Lengua de Señas mediante un software traductor sería un gran aporte a la sociedad, permitiría suplir en cierta medida el rol de los intérpretes y abriría puertas a aquellos sordos y/o mudos que deseen superarse.
Si bien aún estamos lejos de obtener un resultado completo y comercialmente abierto al público, cada estudio y propuesta realizada por terceros apoya invaluablemente a la comunidad de interesados en acercarnos un poco más a la solución final tan ansiada. [4]

1.9 Alcance

La presente tesis tiene como finalidad establecer las bases para la obtención de un Sistema Traductor de Gestos Dinámicos. Con la intención de conseguir lo anterior es necesaria la ejecución de una serie de pasos y definición de un alcance que permita dar a conocer claramente lo que se planea lograr.

- La solución propuesta sólo se enfocará en desplazamientos de brazos y manos, se obviarán aspectos como expresiones faciales, posturas de manos y torso.
- Se realizará una revisión de los métodos y técnicas empleados en general a lo largo de los últimos años que hayan sido propuestos como solución al problema y similares. Así también una evaluación y sustentación del por qué se elige al sensor Kinect como herramienta de trabajo.
- Se realizará el estado del arte de los modelos y técnicas que emplean la Kinect como herramienta para la solución.
- Se definirá un conjunto de gestos a reconocer.

- Se desarrollará un repositorio software o base de datos donde estarán registrados los datos correspondientes a cada gesto con el que se piense trabajar.
- Se empleará el sensor Kinect para la obtención de datos a partir de escenas en tiempo real.
- Se desarrollará un software que realice el procesamiento de datos a través de las técnicas y métodos seleccionados con la finalidad de obtener la traducción o resultado esperado.
- Se planteará la ejecución del sistema en tiempo real, empleando datos obtenidos mediante la conexión física al sensor Kinect.
- Se diseñará y desarrollará una interfaz gráfica que permita visualizar las capturas en tiempo real y los resultados del procesamiento de los datos obtenidos.
- Se realizarán pruebas empleando el conjunto de señas previamente establecido.
- Se ejecutarán validaciones contrastando tasas de reconocimiento y costos computacionales con trabajos de terceros y causantes de diferencia.

CAPÍTULO II: MARCO TEÓRICO

Para el desarrollo de este trabajo es importante entender la naturaleza de los gestos humanos desde diversos aspectos. Esta sección tiene por finalidad empapar al lector con los conceptos más importantes e ideas centrales que influyen en el desarrollo y diseño del producto final buscado.

2.1 Gestos

Un Gesto es conocido como cualquier movimiento realizado por el rostro, manos u otras partes del cuerpo, con la intención de transmitir mensajes, ideas y/o expresar afectos. El Gesto engloba la representación de un carácter mediante el empleo de movimientos. Mientras el término Postura referencia la simbolización de un carácter a través de las manos con una configuración estática, es decir sin movimientos ni trayectorias realizadas.

MacNeill en [11] busca reconocer y evidenciar si existe relación entre los gestos y el lenguaje hablado. El trabajo empieza definiendo Gestos como movimientos corporales equivalentes a algún significado y cuya performance es con cierto fin. Con dicha denominación los Gestos quedan encasillados a un conjunto de acciones corporales realizadas de forma visible y voluntariamente por parte del emisor. En contraste, el autor da cuenta de otras fuentes las cuales la definen como acompañantes del diálogo hablado y consecuencia directa de convenciones sociales.

Antes de postular una definición propia, el autor considera evaluarlo desde un enfoque de participación de los gestos tal y como se da en las interacciones habladas, partiendo de la existencia de similitud entre el Lenguaje Hablado y el Gesto al ser usados como medios para representar algo, y considerando además que los gestos emplean otro camino para lograr la simbolización.

Mediante los gestos se puede lograr la representación de objetos, concretos o metafóricos, a través de la creación y empleo de trayectorias o movimientos. Por otro lado, los gestos típicamente son usados como medio para añadir "Capas de significado" a lo que se está hablando, esta particularidad permite sobrepasar el valor de lo expresado hasta cierto punto. Es decir, los gestos bajo el rol de acompañante pueden influenciar considerablemente la interpretación de los receptores de lo emitido por el orador. Así

también, el empleo de gestos permite al emisor ser más preciso en lo hablado o dar un significado y sentido más completo.

Con la intención de evidenciar la relación entre el Lenguaje y los Gestos, el autor postula dos enunciados, proponiendo casos y ejemplos concisos en cada uno:

1) Los gestos cumplen la función de otorgar contexto a las expresiones verbales habladas.

2) Los gestos permiten aportar contenido sustantivo adicional.

Para el primero el autor plasma el resultado de una evaluación consistente en pedir que alguien narre lo mejor posible una fábula a un grupo de niños de un colegio. En esta evaluación se concluyó que las mismas palabras/verbos usados pueden estar acompañados de diferentes gestos, dependiendo directamente del contexto el cual se quiera generar sobre los oyentes.

Para el segundo enunciado, el autor reconoce como característica común que los oradores realizan gestos durante su diálogo e incluso los mantienen después de haber terminado de hablar (Por ejemplo, se observó a ciertos emisores sostener sus manos aun formando el último gesto) a pesar de que otra persona ya se pueda encontrar hablando. Esta gesticulación ayuda al orador a mantener y evidenciar mediante una muestra su estado de ánimo o dar a notar la intención de su última expresión.

Finalmente, el autor resalta que los gestos y el diálogo hablado empleados en una conversación son básicamente diferentes, pero tienen roles complementarios. Dado ello, no considera necesario que los gestos logren características del lenguaje hablado mientras esté disponible en las facultades de la persona.

En contraste con la definición anterior, England [12] toma una postura más técnica al referirse a la realización de gestos enfocados en la implementación de una interface de interacción basada en gestos.

Para postular una definición, el autor empieza referenciando las ideas de algunos autores, tales como la establecida por Kendon, quien explica en términos generales que los Gestos son expresiones y movimientos voluntarios del cuerpo realizados durante el diálogo y son considerados por los receptores como parte del significado de la conversación. Esta definición excluye explícitamente los factores externos (Contexto y el diálogo el cual se esté realizando) que pueden influir y ayudar a determinar si cierta acción es un gesto. Sin

embargo, Kendon no propone una forma para reconocer los gestos. Como consecuencia, el autor considera es una definición incompleta.

Otra fuente referenciada es Cassell, cuya definición limita los Gestos a movimientos realizados por las manos durante el diálogo. Cassell reconoce que ello no basta para abarcar el alcance de los Gestos, su definición se basa en que muchos sistemas de reconocimiento de gestos se centran sólo en "Lenguajes Gestuales" no teniendo en cuenta los movimientos que se dan cuando se realizan diálogos, siendo necesario considerarlos al crear interfaces de interacción basadas en reconocimiento de gestos.

Finalmente se cita la definición brindada por Väänänen and Böhm, quien define al Gesto como cualquier movimiento del cuerpo empleado para transmitir alguna información a otra persona. Este último es considerado por el autor como muy general pero capaz de abarcar el verdadero significado.

Las interfaces de interacción basadas en gestos no han sido ampliamente empleadas hasta hace algunos años cuando herramientas cotidianas como los celulares comenzaron a venir equipados con sensores como acelerómetros. Estas nuevas características técnicas facilitaron el reconocimiento de gestos para algunas aplicaciones en particular en lo que ahora se conoce como smartphones.

El autor reconoce la necesidad de estudios previos para determinar qué gestos serían socialmente aceptables por futuros usuarios de las aplicaciones software para ser realizados y reconocidos en ambientes públicos.

Varios autores reconocen diferencias entre los gestos empleados normalmente al dialogar y los usados en interfaces de interacción basados en gestos, además reconocen que estos últimos requieren un establecimiento previo de los gestos que deberán ser reconocidos, por lo cual no cabrían correctamente en la definición que se maneja de Gestos. No obstante, el autor en su trabajo también hace hincapié a que no todos los gestos son fácilmente aceptados y realizados en espacios públicos por diversos factores como el contexto cultural, esto último es un criterio que se debe tener en cuenta al definir los gestos.

Como ha podido observarse, los Gestos pueden entenderse como movimientos realizados con la intención de comunicar algo, para lograr ello no solamente pueden emplearse movimientos de manos y brazos, sino además el uso de posturas, desplazamientos de cabeza, miradas, expresiones faciales, entre otros. Se conoce además que los gestos han

sido estudiados por diversas disciplinas (Tales como: La psicología, antropología, comunicación, danza, expresiones corporales, lingüística y ciencias de la computación) proponiendo definiciones particulares a partir de diversos criterios. En el campo de la Informática estas definiciones se logran entrelazar y mezclar, definiendo a los Gestos como todo movimiento llevado a cabo mediante la implicancia de manos, brazos, cabeza, mirada y las expresiones faciales.

En ese bagaje de definiciones y estudios sobre los gestos, existen autores quienes se enfocaron en investigar y proponer clasificaciones bajo diferentes criterios y puntos de vista.

2.1.1 Clasificación de los Gestos

Nos hemos basado en investigaciones de terceros con la finalidad de dar a conocer los tipos de gestos con los cuales se trabajará. Estos han sido clasificados bajo aspectos como sus propiedades y relaciones con el habla, pensamiento, así como sus características matemáticas y lingüísticas.

2.1.1.1 Clasificación de McNeill

McNeill [13] propone una clasificación con la intención de estudiar los gestos empleados cuando las personas intentan transmitir acontecimientos o sucesos que acaban de percibir empleando solamente gestos. Considerando ese enfoque y otras experiencias McNeill reconoció cuatro tipos principales:

- Gestos Icónicos, en esta categoría se encuentran los movimientos capaces de otorgar información de las características físicas o relaciones espaciales reconocidas de los objetos, tales como el tamaño y la forma. Por ejemplo: Una torta sobre la mesa es comúnmente representada por una mano girando en círculos con el dedo índice apuntando hacia abajo. Una mano subiendo puede representar la elevación de algún objeto o alguien. Este tipo de gestos son recurrentemente empleados en la Interacción Humano Computadora, unos grandes ejemplos son los realizados al acercar o alejar el zoom, también al representar letras, golpes y movimientos.

- Gestos Metafóricos, usados para representar imágenes abstractas e ideas. Generalmente cuando alguien tiene una pregunta extiende la mano en forma de copa esperando recibir una respuesta de alguien.
- Gestos Ilustrativos, acompañan la comunicación verbal con la finalidad de dar énfasis a lo que se menciona en forma rítmica, pueden reemplazar palabras en situaciones difíciles para el orador. Se consideran carentes de contenido semántico de relevancia.
- Gestos Deícticos, empleados para señalar y/o seleccionar objetos, así también para indicar direcciones. En la Interacción Humano Computadora resalta este tipo de gesto gracias a su carácter espacial (Atrás, adelante, entrar, salir, moverse, etc.).

2.1.1.2 Clasificación para su Reconocimiento

Dentro del ámbito de la Interacción Humano Computadora (IHC), se puede considerar una clasificación de gestos dependiendo de si son Estáticos o Dinámicos, incluyen una o varias partes del cuerpo, Faciales o Corporales, consideración de data 2D o 3D, y Unimodales o Multimodales.

- Gestos Dinámicos versus Gestos Estáticos

  Los Primeros comprenden movimientos de una o varias partes del cuerpo, su reconocimiento implica modelado en el espacio, así como la segmentación temporal donde se haya reconocido e identificado correctamente el comienzo y la finalización del gesto.

  Los Segundos son configuraciones específicas y estáticas de la estructura del cuerpo, por dicha razón también son llamados Poses.

- Una Parte del Cuerpo versus Varias Partes del Cuerpo

  Los Primeros son estructuras estáticas a las cuales se debe ajustar una determinada parte del cuerpo, como si se realizara una postura.

  Los Segundos implican considerar aspectos como la sincronización entre diferentes partes del cuerpo comprometidas, además para determinados casos algunas partes del cuerpo pueden empezar su trayectoria antes o después de otras, elevando la complejidad.

- Gestos Faciales versus Gestos Corporales

  Los Gestos Faciales conciernen los movimientos producidos por los ojos, cejas y labios. Por su parte, los Corporales involucran movimientos de distintas partes del cuerpo tales como: Los pies, brazos, manos, cabeza, etc.

- Reconocimiento 2D versus Reconocimiento 3D

  El reconocimiento de gestos idealmente debe realizarse con datos 3D. No obstante, algunos autores han considerado realizar el trabajo no incluyendo información de profundidad (Reconocimiento 2D), obligándolos a superar factores externos (Tales como la eliminación del fondo y la iluminación), empleando mayor cantidad de recursos.

- Sistemas Unimodal versus Sistemas Multimodal

  Ciertos sistemas de reconocimiento emplean sólo una fuente de datos de entrada (Unimodal) para realizar la tarea, mientras otros autores más diestros y avezados en la búsqueda de mayor realismo y exactitud se han basado en una serie de inputs que pueden variar de naturaleza o tipo, tales como: Audio, imágenes de profundidad, imágenes RGB, acelerómetros, entre otros.

Como puede observarse, las tipificaciones existentes permiten deducir la cantidad de requerimientos necesarios para el modelado y proceso de reconocimiento según cada caso.

En términos de la Clasificación de McNeill el tipo de reconocimiento a realizar en este trabajo estará enfocado en Gestos Icónicos. Además, se trabajará sobre el reconocimiento de Gestos Dinámicos y Corporales, considerando solamente el desplazamiento de brazos, codos y manos (No posturas de las manos, ni dedos), empleando data obtenida únicamente por un sensor Kinect (Unimodal) para el reconocimiento 3D, mayor detalle se da en el apartado nombrado Alcance.

2.2 Interacción Humano Computadora

Ghaoui [14] menciona a La Asociación de Maquinaria Informática (ACM – Association for Computer Machinery), esta es una organización cuyos participantes (Investigadores y profesionales en general) se interesan en el área de las Ciencias de la Computación. Dentro de esta entidad existen grupos tales como el SIGCHI (Special Interest Group in Computer Human Interaction), enfocado en temas de Interacción Humano Computadora, el cual reconoce que HCI ha venido influenciando fuertemente el rumbo de la tecnología y su uso en nuestra vida cotidiana, para formalizar ello reconocen que actualmente no existe una definición capaz de englobar el número de áreas y aspectos involucrados, sin embargo ofrecen un concepto base: "Disciplina relacionada con el diseño, la evaluación e implementación de sistemas informáticos interactivos para el uso de los seres humanos, y con el estudio de los fenómenos más importantes con los que está relacionado". Para la perspectiva de las Ciencias de la Computación está centrada en la interacción entre uno o más humanos con una o más computadoras mediante diversos medios (Como programas de interacción gráficos). Debido al gran bagaje de disciplinas en las cuales puede aplicarse, el HCI es considerado multidisciplinario, abarcando la Psicología, Ciencias Cognitivas, Ergonomía, Sociología, Negocios, Diseño Gráfico, Informática y Desarrollo de Software, entre otros.

Según Jounghyun [15], desde los últimos años HCI ha cobrado mayor relevancia en varios aspectos de nuestras vidas debido a su naturaleza multidisciplinaria abarcando principalmente áreas como Ingeniería, Psicología, Ergonomía, entre otros. Se enfoca no sólo en la funcionalidad, sino también en la usabilidad buscando lograr facilidad de uso en las interfaces para permitir al usuario realizar tareas de forma eficiente y eficaz, intuitivamente sin requerir mayor entrenamiento con la interface. Por lo anterior se dice que la HCI permite la evaluación de diferentes formas por las cuales se puede dar la interacción entre humanos y dispositivos computarizados, influenciando conceptos, diseños, implementaciones y evaluaciones dadas durante el desarrollo.

Esta fuente resalta que la palabra "Interaction" de "Human Computer Interaction (HCI)" hace referencia a dos conceptos: Interacción e Interface. La primera debe entenderse como un modelo abstracto mediante el cual los humanos interactúan con los dispositivos computarizados con alguna finalidad. Mientras Interface referencia al hardware o software producto de la adopción de algún modelo abstracto de interacción.

La Interacción Humano Computadora (IHC) cuenta con la característica de copar diferentes secciones de investigación, entre ellas el Reconocimiento de Voz y el Reconocimiento de Lengua de Señas son sólo un par del total. El Primero ha contado con notable progreso y especial auge en las últimas décadas, no obstante, a pesar del esfuerzo de investigadores y terceros se puede decir que aún se encuentra en etapas de desarrollo. Esta lleva por objetivo el control de las computadoras o herramientas tecnológicas en general mediante el empleo de comandos de voz. Por su lado, el Reconocimiento de Lengua de Señas ha sido un área de investigación en crecimiento también atractiva para los investigadores y desarrolladores en general, especialmente desde que sensores de profundidad como la Kinect fueron puestos en venta a precios accesibles. Sin embargo, la investigación y resultados obtenidos en este rubro aún no alcanzan el nivel de madurez deseado.

2.3 Sensores

Básicamente se definen como dispositivos creados con la finalidad de capturar o detectar cambios (Datos analógicos) y convertirlos en datos a ser enviados (Datos digitales) para su posterior procesamiento.

Según Juran [16], los Sensores son dispositivos detectores especializados. La finalidad de su creación y diseño es reconocer y verificar la presencia e intensidad de un fenómeno para posteriormente entregar la data obtenida. Dichos valores deberían ser procesados por el receptor con la intención de tomar decisiones. El término “Sensor” engloba a todos los instrumentos técnicos al igual que los sentidos de los humanos y animales, es decir, se denomina así a toda cosa que puede percibir algo.

Según Pallás [17], el término Sensor es empleado de forma técnica para describir a los dispositivos que evaluando la energía del medio o contexto en el cual se encuentran, emiten señales de salida de forma proporcional a la variable medida. Es considerado un medio para obtener conocimiento de cantidades físicas los cuales no pueden ser detectados directamente por los sentidos debido a su naturaleza o tamaño.

Para Vetelino [18], el actual surgimiento y desarrollo de los sensores es consecuencia de años de avances en campos correspondientes a la tecnología e ingeniería. Los sensores tienen la característica de traducir fenómenos físicos en salidas eléctricas con la finalidad de ser procesadas por dispositivos computarizados.

El protagonismo de los sensores en diversas áreas es producto de atributos como: La constante mejora y desarrollo de sus características buscando ofrecer mayor fidelidad en los datos entregados a un costo proporcional a la performance.

En dicho artículo se explica la existencia de ambigüedades en las definiciones de tecnologías relacionadas con los sensores como secuela de contar con un rol primordial en diferentes disciplinas de constantes cambios y desarrollo. Ello imposibilita establecer una definición unánime capaz de englobar la totalidad de áreas. Los autores reconocen que los investigadores consideran diferentes las definiciones de Sensor y Transductor. Según ANSI (The American National Standards Institute), un Transductor es "Un dispositivo que proporciona una salida utilizable como respuesta a una medición específica", sin embargo, la diferencia con el Sensor radica en que este último es ampliamente empleado en la literatura científica por lo cual el autor considera emplearlo como término también.

Dada la definición e inexistencia de precisión en los componentes físicos que forman parte de los sensores, en el libro se proponen dos términos básicos:

- ✓ "Elemento Sensor: Es el mecanismo de transducción fundamental que convierte una forma de energía en otra. Algunos sensores pueden incorporar más de un Elemento Sensor."
- ✓ "Sensor: Un Elemento Sensor el cual incluye su embalaje físico y conexiones externas."

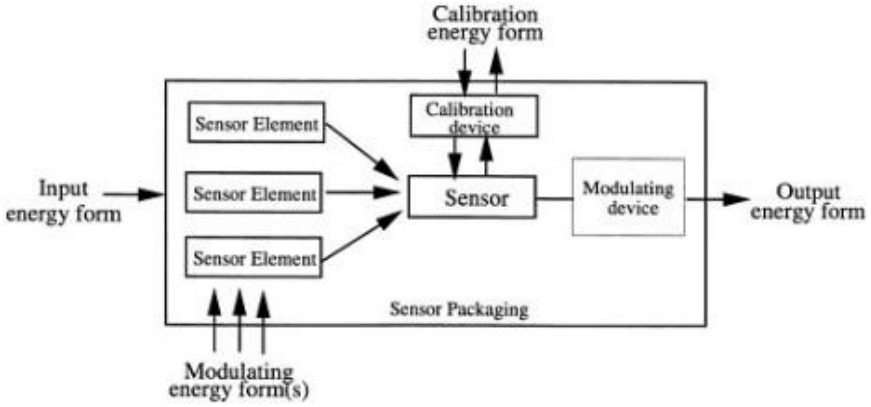


Figura 2.1 Anatomía de un sensor. Fuente: [18]

Durante los últimos años se han dado diferentes estudios y propuestas de clasificación, algunas basadas en el tipo de input esperado por el sensor. Así por ejemplo existen según:

- El principio físico o químico de transducción.
- El tipo de variable de entrada primaria.
- El tipo de aplicación del sensor.
- El costo.
- La exactitud.
- El tipo de output.

En los últimos 10 años la tecnología de sensores se ha centrado en el desarrollo de Sensores Inteligentes, estos difieren de los clásicos al manejar complejidades internamente volviéndolas transparentes para el sistema host (Receptor), presentando una "Cara simple" mediante una interfaz digital. Es decir, la complejidad es soportada por el sensor y no por el sistema de procesamiento de la data entregada.

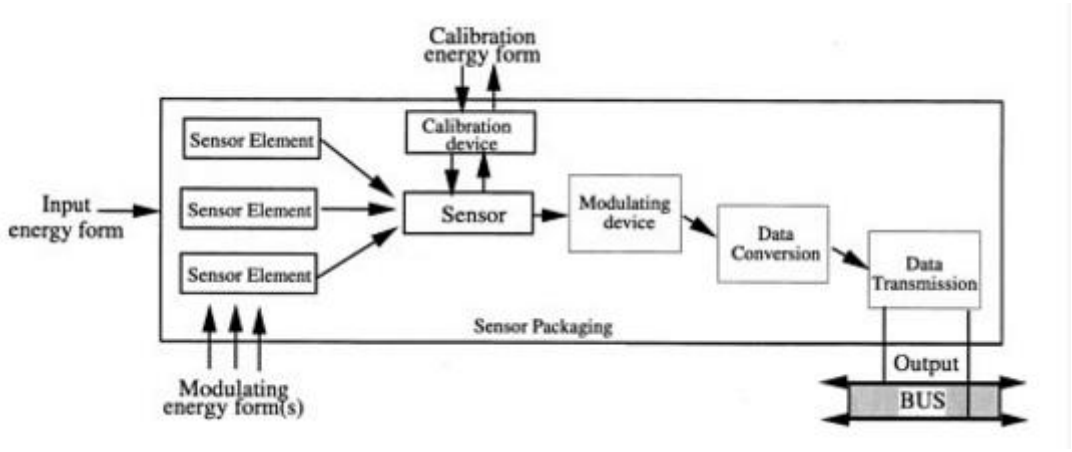


Figura 2.2 Representación de un sensor inteligente. Fuente: [18]

Según Wilson [19], un sensor es un aparato con la posibilidad de convertir fenómenos o variaciones físicas en señales eléctricas, siendo mediador entre el mundo físico y el eléctrico (Por ejemplo: Dispositivos computarizados).

Los sensores han recibido gran impulso en su desarrollo durante los últimos años debido a su capacidad y aplicabilidad en diferentes campos y productos. Desde que estos dispositivos han contado como output señales eléctricas, han adquirido características de dispositivos electrónicos, ello se ve reflejado en la semejanza de sus datasheets con los de artefactos electrónicos.

La existencia de dichos datasheets y la ausencia de algún formato o estándar internacional conducen a una constante confusión en la definición de términos relacionados a los sensores.

Los Datasheets son documentos que permiten al patrocinador resaltar los aspectos positivos y usos potenciales del sensor, funcionando como medio de marketing.

Entre las características del sensor, se pueden mencionar:

- Sensibilidad: En general, es la variación de la Señal de Salida que resulta de un cambio en el Input. Es decir, un mínimo cambio en la Señal de Entrada debe reflejarse en un cambio proporcional en la Señal de Salida. Entre mayor sensibilidad, mayor será la exactitud de la medición resultado.
- Alcance o Rango Dinámico: Es el rango admisible como Señal de Entrada (Input) que puede ser convertida en Salida (Output) sin problemas por parte del sensor. Probablemente se obtenga una gran imprecisión como resultado, o daños al sensor si la entrada está fuera del rango.
- Incertidumbre: Debe entenderse como el mayor error que se puede obtener al comparar la señal de salida real y la ideal.

2.3.1 Sensores 3D

Acorde a Li [20], la característica principal del Sensor 3D es permitir la estimación de la información de profundidad, basándose en diferentes técnicas para obtener los valores entre el sensor y los objetos en escena. Dicha información puede funcionar como discriminatoria en diferentes aplicaciones al ser invariante ante cambios de iluminación o el pigmento de la piel, superando así a las capturas de cámaras 2D convencionales. En este trabajo se direcciona el empleo de los Sensores 3D para el Reconocimiento de Gestos.

Los mecanismos de detección son variados, siendo el común denominador proporcionar imágenes de color e información de profundidad para cada pixel por rango de tiempo (Normalmente cada segundo). Un buen ejemplo de este tipo de tecnología comercial ampliamente usado es Swiss Ranger cuyo costo se aproxima a los 10.000 dólares, resultando notablemente caro.

La variedad parte de la existencia de distintos tipos de sensores y cámaras con posibilidad de medir la profundidad, entre estos se encuentran las cámaras de Visión Estereoscópica[9] y las basadas en técnicas de Rango de Imagen[10] (Como los Sensores Time-of-Flight y Luz Estructurada). El tipo de sensor a elegir para la realización de un trabajo o proyecto debe resultar del análisis del caso y requerimientos a atender, ya que cada sensor entrega distintos niveles de precisión en las medidas de profundidad, resolución, frecuencia de captura, entre otros.

Bajo esta misma clasificación también se encuentran aquellos caracterizados por depender de interfaces Wearables[11], singularizados por ser parte de la vestimenta del usuario para obtener data. En este rubro se encuentran las Google Glass (Disponen de una cámara con la cual reconocen los movimientos de brazos y manos), Samsung's Galaxy Gear smartwatches, Pebble smartwatches, entre otros.

2.4 Cámaras RGB-D

Peter [21] define a las Cámaras RGB-D como capaces de capturar imágenes RGB convencionales de forma conjunta con información de profundidad por pixel a frecuencias aceptables. Estos dispositivos tienen décadas de historia, siendo especialmente desarrollados y empleados en grupos de investigación interesados en la visión por ordenador. Sin embargo, en los últimos años han recibido un gran impulso proveniente de empresas dedicadas al desarrollo de juegos por ordenador y entretenimiento doméstico en general.

---

[9] Infieren información de una escena 3D a partir de imágenes tomadas desde distintas posiciones al mismo tiempo. Se asemeja a la visión binocular realizada por las personas.

[10] Producen imágenes 2D que muestran la distancia a los objetos en escena desde un punto específico. Generalmente son logrados mediante la participación de algún sensor.

[11] Son aquellos dispositivos que se llevan sobre, debajo o incluidos en la ropa, generalmente están siempre encendidos. Puede trabajar como multitarea por lo que no requiere dejar de realizar algo para usar el dispositivo, pudiendo actuar como extensión del cuerpo. Actualmente se emplean en las áreas de: Salud, deporte y bienestar, entretenimiento, industrial y militar.

Figura 2.3 Imagen RGB (Izquierda) e imagen de profundidad (Derecha) capturados por una cámara RGB-D. Fuente: [21]

Para lograr las estimaciones de profundidad considerando el gran número de pixeles que deben manejar según la resolución de las imágenes, las Cámaras RGB-D se basan principalmente en el empleo de dos tipos de tecnologías: Luz Estructurada y Time-of-Flight Sensing (Detección de Tiempo de Vuelo).

No obstante, estos dispositivos muestran limitaciones en cuanto a cantidad de información capturada en comparación con sistemas especiales cartográficos 3D, originados por diferencias en las características técnicas. Entre las diferencias citadas el artículo menciona que el rango de captación de la información de profundidad es por lo general hasta cinco metros y con un campo de visión de 60º, mientras los sistemas especializados cuentan con un rango de 180º y proporcionan mucho menor ruido en la data entregada.

Lograr la construcción de mapas de entornos 3D es tarea importante en trabajos de navegación, robótica, manipulación, telepresencia, realidad aumentada, entretenimiento y otros. Enfoques basados sólo en el empleo de Nubes de Puntos 3D son muy aceptables para la reconstrucción 3D, pero corren el riesgo de ignorar información importante presente en imágenes de color. Por otro lado, los sistemas especializados de mapeo 3D resultan costosos, lentos en el procesamiento y entrega de información. Por su parte, los enfoques que emplean sólo imágenes 2D para la reconstrucción 3D requieren mucho cálculo computacional, pierden robustez y carecen de la madurez necesaria para entregar modelos de densidad 3D a precisión aceptable. En estos últimos, el mayor o menor ruido es consecuencia directa de la presencia de factores externos en el ambiente donde se realiza la captura. La limitada iluminación, la imposibilidad de reconocer y diferenciar características, y la demanda de mayor precisión (Información) son unos cuantos factores

a tener en cuenta al volver la tarea de reconocimiento más difícil al realizar aplicaciones de este tipo.

Para Bouwmans [22], las cámaras RGB-D son una nueva generación de dispositivos capaces de captar y permitir emplear data de profundidad para lograr identificar y rastrear segmentos de partes humanas. Estos sensores recibieron gran impulso por investigaciones en el área de Visión por Computador, entre estos se encuentran aplicaciones dirigidas al monitoreo del cuidado de salud en el hogar y en aplicaciones de robots.

Estos artefactos inicialmente fueron presentados como controladores libres para juegos buscando romper el paradigma del uso obligatorio de mandos para la interacción, con el paso del tiempo se entendió el potencial de los mismos en diversas aplicaciones de ayuda social empezando a ser empleados para el análisis de movimientos corporales más complejos y con diferentes fines.

El creciente interés de investigadores en el uso de estos dispositivos es consecuencia directa de dos factores: La calidad de Performance (Frames por segundo y la Resolución de imagen) entregada a precio accesible, y que la Data de Profundidad resulta atractiva y adecuada en el desarrollo de aplicaciones. Este tipo de data no se ve afectada por problemas típicos en el reconocimiento basado únicamente en imágenes de color los cuales impactan seriamente en la eficiencia y eficacia del aplicativo. Entre dichos problemas se encuentran: Fragmentación de imagen y fondo, incorrecta diferenciación de regiones que cuentan con la misma tonalidad de color, inconvenientes de identificación de forma gracias a las sombras generadas por objetos/imágenes ubicadas en primer plano, problemas causados por modificaciones en la iluminación del ambiente, entre otros.

Ling [23] los describe como sistemas físicos de detección que capturan imágenes RGB con información de profundidad por píxel. Para estimar los pixeles de profundidad se basan principalmente en dos tipos de tecnologías: Estéreo Activo o detectores Time-of-Flight. Este tipo tecnología existe hace décadas, usados ampliamente en diversas áreas mediante aplicaciones de visión por computador debido a sus bajos costos, mayor exactitud y robustez en los datos entregados en comparación con los obtenidos por cámaras 2D. Sin embargo, los métodos empleados (Algoritmos) aún deben tratar con problemas generados por el ruido al capturar datos de profundidad.

Permiten capturar data de profundidad a exactitud razonable y a una alta frecuencia de datos. Sin embargo, estas cámaras cuentan con algunos inconvenientes con respecto al

mapeo 3D: Proveen datos de profundidad correctamente desde y hasta cierta distancia (Generalmente menos de 5 metros), dependiendo de las características técnicas del dispositivo sus estimaciones de profundidad tienden a ser ruidosas y el campo de visión es aproximadamente de 60 grados, siendo limitada en comparación con cámaras especializadas y escáneres láser empleados para cartografía 3D (180º aprox.).

En este ámbito, el mercado es liderado por Microsoft Kinect y Leap Motion Controller debido a sus costos accesibles. Ambos han sido usados exitosamente en diversas áreas logrando una aceptable tasa de reconocimiento, una mayor aprobación al ser instrumentos menos intrusivos e idóneos para ser empleados en entornos reales.

Las Imágenes RGB-D están compuestas por dos imágenes 2D: La primera es una imagen RGB convencional y la segunda es de profundidad. La Imagen RGB lleva por finalidad otorgar información de la textura del mundo enfocado por el sensor, mientras la Imagen de Profundidad añade geometría, luego gracias a los parámetros de calibración de la cámara (Centro óptico y distancia focal) permite generar Nubes de Puntos 3D. Una Nube de Puntos es definida como una colección de puntos tridimensionales. Ésta tiene la posibilidad de ser generada de forma artificial o provenir de capturas de elementos del mundo real. Las coordenadas $\{p_x, p_y, p_z\}$ de cualquier punto p que compone la nube están dadas con referencia a un sistema de coordenadas fijo (Origen). Si la nube representa datos del mundo real entonces el origen del sistema de coordenadas suele ser el sistema o dispositivo de captura empleado. En este caso, el valor de cada punto p representa la distancia desde el origen hasta la superficie de los objetos en escena donde el punto fue capturado. Dependiendo de las características técnicas del sensor, las nubes de puntos pueden incluir mucha más información que sólo posiciones 3D, es decir, también puede contener información de los puntos como el color, la intensidad, puntos de vista, entre otros.

2.5 Kinect

Según Salazar [24], en junio del 2009 bajo el nombre de "Proyecto Natal", Microsoft anunció el surgimiento de una nueva línea de controladores físicos económicamente accesibles enfocados en permitir la interacción con los videojuegos de su consola Xbox 360 mediante detección de gesticulaciones realizadas por el cuerpo humano, como respuesta al PlayStation de SONY y al Wii de Nintendo.

El proyecto tuvo origen en Brasil, siendo liderado por Alex Kipman. Surgió con el propósito de facilitar la interacción con los videojuegos mediante el empleo de gestos y voz buscando romper el paradigma tradicional del empleo de controladores físicos o mandos. El nuevo paradigma se basó en sensores acústicos y visuales, permitiendo a los usuarios interactuar con los juegos electrónicos mediante gestos y comandos de voz.

La primera versión oficial del sensor Kinect se publicó el 04 de Noviembre del 2010, siendo sólo compatible con la Xbox 360, una consola de videojuegos desarrollada por Microsoft. Una vez que se reconoció su verdadero potencial, Microsoft lanzó una versión beta del dispositivo enfocado para desarrolladores en Windows. Dicha versión salió al mercado el 16 de Junio del 2011, siendo inicialmente compatible con Windows 7 y los lenguajes de programación C++, C# y Visual Basic .NET.

Posteriormente, Microsoft liberó un Software Development Kit (SDK) brindando la posibilidad de aprovechar al máximo la data provista por el sensor en el desarrollo de aplicaciones basadas en Kinect en lenguajes como C++, C# y Visual Basic.

Este dispositivo ha vendido más de 24 millones de unidades a nivel mundial y ganado premios como invento innovador.

Según Sarbolandi [25], la Kinect tiene la posibilidad de “Ver” la escena en tres dimensiones y hacer el seguimiento en tiempo real. El dispositivo obtiene datos de la cámara y del sensor con el cual cuenta. Una vez obtenida la data, esta es procesada por el software (SDK) para “Interpretar” la escena, detectar personas y rastrear los movimientos del esqueleto de cada una (Skeleton Tracking), pudiendo reconocer hasta 2 (Kinect versión 1.0) y 6 (Kinect versión 2.0) en escena. La Kinect provee capturas tridimensionales de cada cuerpo humano reconocido, formando un esqueleto virtual (Figura 2.4) resultante de conectar el conjunto de puntos 3D correspondientes a las articulaciones (Joints) del sujeto rastreado. A partir de este procesamiento y abstracción

es que desarrolladores e investigadores han empleado los datos obtenidos por el sensor de forma conjunta con métodos algorítmicos para lograr detectar y reconocer los gestos realizados por individuos mediante la comparación de los trayectos recorridos por los joints frente a un patrón ya definido.

Dicho invento fue una revolución al permitir el reconocimiento de gestos, facial y de voz a precio bajo en comparación con otros sensores, convirtiéndose en una opción accesible para el desarrollo de Interacción Humano Computadora (HCI) en tiempo real.

Con estas ventajas los investigadores comenzaron a desarrollar gran cantidad de aplicaciones innovadoras multidisciplinarias (Cuidado de la salud, realidad aumentada, entre otros).

Recientemente Microsoft publicó una segunda versión de la Kinect enfocada en su consola de juego Xbox One, basándose en el principio Time-of-Flight, este dispositivo cuenta con características superiores a su antecesor en tanto a precisión en reconocimiento de partes humanas, como número de personas en escena.

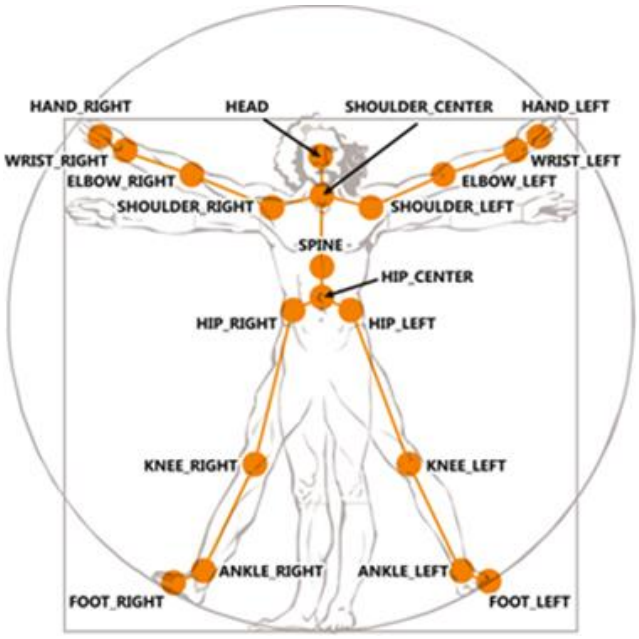


Figura 2.4 Joints reconocidos en el skeleton tracking del SDK 1.8 de la Kinect. Fuente: [26]

2.5.1 Ventajas y Desventajas

El sensor Kinect cuenta con algunas ventajas que lo diferencian de otros del mismo rubro:

a) Reconocimiento de voz y reconocimiento facial. El sensor Kinect está habilitado para recibir comandos u órdenes de voz en diferentes idiomas. Además, soporta el reconocimiento facial y seguimiento del rostro.

b) Skeleton Tracking. Es decir, la representación del esqueleto humano en escena puede ser realizado mediante el seguimiento y detección de las articulaciones.

c) Instalación sencilla. El sensor Kinect cuenta con un SDK que puede ser descargado e instalado en el sistema operativo Windows. No requiere accesorios adicionales de entrada o conexiones especiales ya que el sensor viene con su propio cable de administración de energía y un cable USB para comunicarse con la computadora.

d) El precio del sensor es aproximadamente 50 dólares, siendo accesible en comparación con otros de su categoría.

e) Motor de inclinación. Permite cambiar la orientación vertical en cualquier ángulo entre +27 y -27 grados.

f) Rango de detección. La Kinect es capaz de reconocer cuerpos humanos desde 0.8 a 4 metros de distancia.

g) Detecta profundidad. Siendo capaz de trabajar en la oscuridad sin rastros de luz.

h) Detecta partes ocluidas. Estima la posición de partes humanas que están fuera del área de detección, basándose en la configuración de los otros joints.

Así también el sensor Kinect presenta algunas desventajas que pueden limitar su capacidad, entre estas se tiene:

a) Trabajar en ambientes externos puede generar problemas en el reconocimiento causadas por la sensibilidad a la luz directa e incluso por recibir luz IR de cualquier fuente.

b) El SDK de la Kinect falla en el reconocimiento cuando se tiene dos o más sensores de su tipo trabajando al mismo tiempo. Las señales infrarrojas pueden interferirse entre ellas.

c) A pesar de la capacidad de la cámara para reconocer rostros y posturas del cuerpo humano, algunas veces puede crear un esqueleto o parte a partir de patas de sillas o mesas en escena.

2.5.2 Áreas de Aplicación

Según Yang [27], la aparición de la Kinect ha tenido gran aceptación en diferentes grupos de investigación alrededor del mundo, reflejado en el desarrollo de aplicaciones multidisciplinarias.

En Marketing se han diseñado aplicaciones como la llamada ARDoor, esta tuvo como fin servir de espejo-probador sobreponiendo ropa en el reflejo del comprador para lucirla sin habérsela puesto.

En el campo de la Medicina se encuentran proyectos importantes de tele-rehabilitación para el tratamiento de Esclerosis Múltiple, un ejemplo es el implementado por la Universidad Rey Juan Carlos. También se tiene TedCas, enfocado en controlar aplicaciones informáticas de quirófano y disminuir el riesgo de infecciones. La Fundación Instituto Valenciano de Neuro-rehabilitación ha desarrollado un sistema que emplea la Kinect para el tratamiento motor y cognitivo de pacientes que han sufrido alguna lesión o enfermedad neurológica. Mientras la ASISPA[12] realizó un ensayo clínico de nueve meses aplicando Kinect para ayudar a enfermos de Alzheimer conseguir una mejor calidad de vida a través de terapias que los mantuviesen activos.

En Negocios se han desarrollado proyectos como el promovido por la Universidad de Carolina del Norte en Chapel Hill, un sistema de videoconferencia 3D el cual empleó cuatro cámaras Kinect.

En Ciencias de la Computación diversas aplicaciones se han diseñado, por ejemplo, el MIT promovió el control de la computadora mediante movimiento de manos y dedos.

Entre otras áreas se tiene: Realidad aumentada, procesamiento de imágenes, herramientas de interacción, reconocimiento de objetos, navegación robótica, educación, tele-presencia, reconocimiento de gestos y reconocimiento de lengua de señas.

[12] Asociación de Servicio Integral y Sectorial para Ancianos

2.5.3 Tecnología Kinect 1.0

Según Borenstein [28], emplear imágenes de profundidad resulta interesante por ser fáciles de entender para el computador en comparación con el uso de imágenes de color convencionales al diferenciar objetos y personas, especialmente si se encuentran en movimiento y bajo constante variación en iluminación. Una imagen de profundidad muestra cuán lejos están los objetos frente a la cámara, permitiendo determinar dónde comienza una y termina otra. Esto hace que el sensor entregue data más confiable para la implementación de aplicaciones software multidisciplinarios.

Pero, ¿cuáles son los componentes físicos que permiten al Kinect trabajar? y ¿cómo logran su objetivo?

Figura 2.5 La Kinect por dentro. De izquierda a derecha: Proyector IR, cámara RGB y cámara IR. Fuente: [28]

Es imposible catalogar al sensor Kinect como una cámara convencional, estas sólo generan imágenes 2D a partir de datos generados por la luz que rebota de los objetos en escena. Por su parte, la Kinect registra la distancia de los objetos que tiene frente, empleando para ello una luz infrarroja a partir de la cual crea una imagen de profundidad cuya finalidad no es conocer cómo lucen los objetos sino aproximar sus ubicaciones.

A primera vista se observan “Tres ojos” en el sensor, ubicados en la parte frontal, dos centradas y el tercero a un lado, este último es parte elemental del funcionamiento al ser el proyector de rayos infrarrojos.

El Proyector de Infrarrojos emite un conjunto de Puntos IR imperceptibles para la visión humana sobre la escena en frente, para luego ser capturados mediante la Cámara IR. En

la siguiente imagen se pueden ver los puntos IR proyectados sobre una superficie plana y un cuaderno.

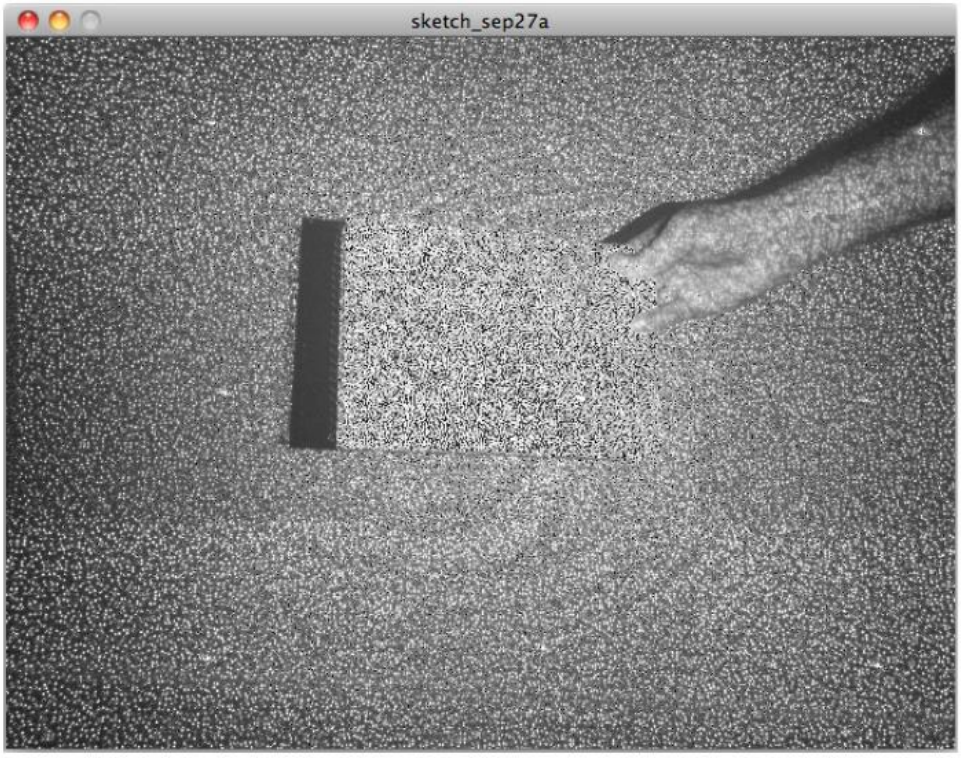


Figura 2.6 Imagen de puntos IR proyectados sobre una superficie y un cuaderno. Fuente: [28]

Antes de obtener los datos se recomienda calibrar el sensor para conocer con mayor precisión dónde se ubican los puntos proyectados en las superficies. Como puede observarse, los puntos sobre el cuaderno tienen una distribución y configuración diferente a los que se encuentran sobre la pared de fondo, es así que, mediante la calibración, el sensor conoce la posición original de todos los puntos y los compara con la distribución de los proyectados en escena. Esta comparación permite conocer cuáles de los puntos están en posición diferente a donde la Kinect espera encontrarlos. Finalmente, la Kinect convierte esta imagen IR en datos de profundidad calculando la distancia de todo lo visible.

Esta forma de trabajo tiene limitaciones, en la imagen anterior puede observarse una sombra detrás del cuaderno, donde ningún punto IR cubre esa zona al verse ocluido por el objeto delante, como consecuencia en dicha sección no se puede determinar la profundidad.

Pero, la detección 3D no es lo único permitido por el sensor, también cuenta con una cámara de color ubicada a un lado de la cámara IR. Esta cámara es similar a cualquier

cámara web con una resolución de 640 x 480 pixeles, su finalidad es proveer la posibilidad de alinear la imagen de color con la de profundidad, de tal forma ser capaz de ocultar imágenes a partir de cierta distancia para mostrar otra a cambio.

Adicionalmente, el sensor cuenta con cuatro micrófonos. Esta cantidad permite no sólo escuchar sonidos en escena, sino también al estar posicionados estratégicamente hacen posible la aproximación de la ubicación del origen de los sonidos. Si se tienen varios jugadores dirigiendo comandos de voz para algún juego, el sensor podrá determinar cuál de ellos está hablando.

Por último, el sensor cuenta con un motor interno cuya finalidad es inclinar de forma automática su estructura y sensores hacia arriba y abajo. Está limitado a 30 grados. Fue implementado para que el sensor se adecúe correctamente a la variedad de escenarios (Salas) según su tamaño, posición de muebles, ubicación de los jugadores, entre otros aspectos.

2.5.3.1 Arquitectura

En Kramer [29] se menciona que el desarrollo de aplicaciones de reconocimiento de gestos basados en el empleo de información de profundidad ha sido capaz de mejorar la interactividad y confort obtenido en los productos implementados, en este aspecto resalta la Kinect.

Físicamente el sensor es un cubo rectangular de aproximadamente 23 centímetros de largo, cuenta con una base hexagonal unida por medio de un eje rotular. El sensor está conformado por una cámara RGB, un conjunto de micrófonos, un sensor de profundidad capaz de capturar el movimiento tridimensional de cuerpos humanos y además posee la capacidad de reconocer la voz y rostro.

En la siguiente imagen se observa de forma referencial la distribución de los principales componentes del sensor:

Figura 2.7 Imagen de la parte externa del sensor Kinect. Fuente: [29]

2.5.3.2 Componentes Hardware

En Kramer [29] se describen los componentes físicos del sensor:

- Emisor Láser IR y Cámara IR. Ambos partes elementales del Sistema de Detección de Profundidad mediante el principio de Luz Estructurada. El primero genera una salida como la mostrada en la Figura 2.6. Mientras la Cámara IR almacena dicho conjunto de puntos reflejados sobre los objetos en escena y los compara con un patrón predefinido de puntos en una superficie plana, el resultado permite reconocer las perturbaciones generadas por las variaciones de las superficies a partir de lo cual se determina la cercanía o lejanía de objetos.
- Cámara RGB. Trabaja a 30 Hz para imágenes de 640 x 512 píxeles y 15 Hz capturas de 1200 x 960 píxeles. Posee un campo de visión de 58 grados horizontales, 45 verticales y 70 diagonales. Trabaja en un área de 0.8 a 3.5 metros.
- Arreglo de micrófonos. Contiene cuatro micrófonos que capturan la información de audio. Su posición hace posible grabar y aproximar las ubicaciones de los orígenes del sonido.
- Motor, permite inclinar entre ±27º adicionales al ángulo de visión vertical de la cámara.
- Acelerómetro, posee un rango de 2G en los tres ejes espaciales donde G es la fuerza de gravitación.
- LED, permite indicar el estado Activo o Inactivo del sensor.

La comunicación entre el dispositivo y la computadora se realiza mediante un USB Hub.

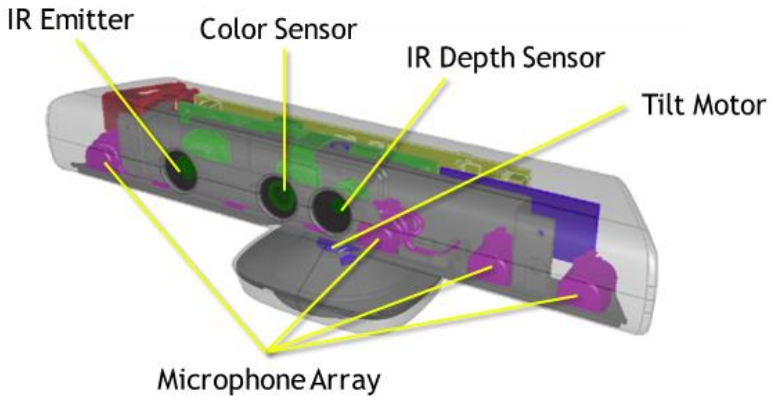


Figura 2.8 Esquema físico del sensor Kinect. Fuente: [29]

2.5.3.3 Especificaciones Técnicas

A continuación, se muestra el resumen de las especificaciones técnicas del sensor Kinect:

Característica	Descripción
Ángulo de Visión	43º vertical 57º horizontal
Rango de Inclinación Vertical	±27º
Frecuencia de Fotogramas (Profundidad y Color)	30 Fotogramas por Segundo (FPS)
Formato de Audio	16-kHz, 24-bit Mono PCM
Características de la Entrada de Audio	Una matriz de cuatro micrófonos con conversor Análogo a Digital de 24 bits (ADC) y procesamiento de señales incluyendo eliminación acústica de eco y supresión de ruido.
Características de Acelerómetro	Un acelerómetro 2G/4G/8G configurado para el rango 2G, con un límite superior a 1º de precisión.

Tabla 2.1 Especificaciones técnicas del sensor Kinect. Fuente: [29]

2.5.4 Controladores

Según Manasrah [30], para el correcto funcionamiento del sensor Kinect es necesario el empleo de un software middleware que facilite la interacción con la computadora, por tal motivo desde la publicación del sensor se han desarrollado una serie de controladores o drivers con la intención de aprovechar la data entregada.

No todos los controladores provienen del mismo fabricante. A continuación, se presenta una comparación entre los dos principales drivers creados, el primero Open Source y el segundo desarrollado por el mismo fabricante (Microsoft).

	SDK OPENNI/NITE	SDK MICROSOFT
Ventajas	Uso libre y comercial.	Soporte para audio.
	Adquiere datos de las manos para seguimiento y reconocimiento de gestos.	Incluye las manos, pies y clavícula.
	Seguimiento de cuerpo completo.	Seguimiento de cuerpo completo.
	Puede calibrar la profundidad y el color de la imagen.	No necesita postura de calibración.
	Cálculo de la rotación de las articulaciones.	Soporte para el motor de inclinación.
	Plataforma múltiple: Windows XP, Vista, 7, Linux y MacOSX.	Mejor tratamiento de articulaciones no visibles.
	Soporte incorporado para grabación y reproducción.	Soporta múltiples sensores.
		Instalación simple.
		Gran cantidad de información disponible.
		Sistema de reconocimiento de gestos.
		La versión 1.7 permite correr el SDK sobre máquinas virtuales.
Desventajas	Instalación compleja.	Licencia única para uso no comercial.
	Sin soporte para audio y motor de inclinación.	Mayor consumo de recursos computacionales.
	Necesita de una postura de calibración.	Solamente para Windows 7 y versiones superiores.

	Las articulaciones no visibles no son estimadas.	Sin soporte incorporado para grabación y reproducción.
	Compatible con varios sensores, aunque la configuración y enumeración es complicada.	Sin soporte para transmitir los datos sin procesar del sensor infrarrojo.

Tabla 2.2 Características de los controladores SDK OpenNI y Microsoft SDK. Fuente: [30]

El OPENNI/NITE es multiplataforma y open software, mientras la desarrollada por Microsoft también es gratis y fácil de obtener, pero mucho más sencilla de instalar. Esta otorga un mejor aprovechamiento de las características y potencial del sensor, siendo capaz de reconocer articulaciones ocluidas y rastrearlas en comparación con OpenNI. Esta ventaja es importante ya que en la realización de cualquier gesto es muy probable la oclusión de algunas partes humanas quedando ocultas a la visión del sensor.

La existencia de mayor documentación, ejemplos y recursos disponibles cobran importancia si se considera el factor tiempo. Se tiene aproximadamente 4 meses para llevar acabo la implementación, pruebas y mejora del software.

Por las razones mencionadas se realizará el desarrollo del presente trabajo empleando el SDK provisto por Microsoft.

2.5.5 El SDK 1.7 de Kinect

Se ha empleado el SDK 1.7 convenientemente en el desarrollo de la presente tesis. Según Iralde [26], reconoce hasta 2 personas en escena, a una tasa máxima de 30 fps las posiciones de 20 articulaciones (Joints) y los ángulos formados por ellas en el espacio virtual tridimensional entre la cámara y los objetos en escena.

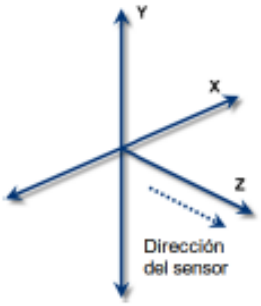


Figura 2.9 Ejes de la Kinect. Fuente: [26]

Para un adecuado reconocimiento el usuario debe posicionarse a una distancia frente la cámara entre 0.8 a 4 metros.

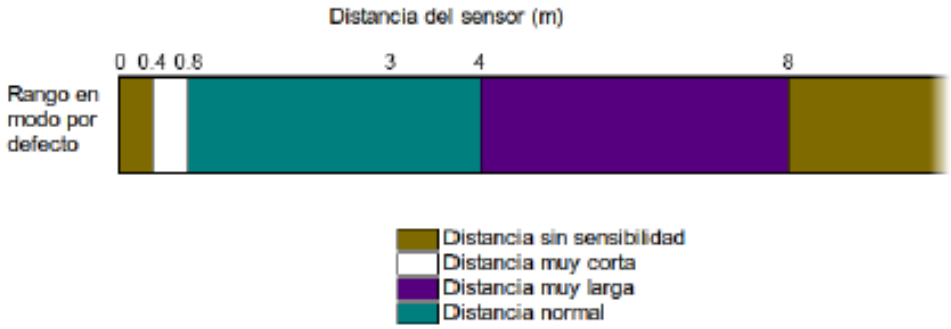


Figura 2.10 Rango de detección del sensor Kinect. Fuente: [26]

El ángulo de visión horizontal es de 57 grados y el vertical de 43 grados. Cuando el usuario se encuentra a 0.8 metros, la cámara captura 87 cm., horizontalmente (Eje x) y 66 centímetros verticalmente (Eje y), estas medidas aumentarán si la distancia a la cámara crece.

2.5.6 El SDK y el Algoritmo Skeletal Tracking

Capilla [31] resalta que el SDK 1.7 del sensor Kinect 1.0 está orientado al reconocimiento de partes gruesas, es decir, en un humano reconocería los movimientos de brazos y manos, más no de los dedos. Además, puede distinguir varias personas y sus movimientos en escena, aun cuando existan oclusiones o no sean visibles para la cámara, gracias a la extrapolación de la posición de las partes ocultas a partir de las visibles.

El algoritmo de reconocimiento de la Kinect trabaja sobre imágenes individuales capturadas por el sensor de profundidad, estas son segmentadas etiquetando partes del cuerpo de forma probabilística; las partes del cuerpo son las ubicadas espacialmente cerca de las articulaciones que se quieren reconocer. Proyectando las partes inferidas en el espacio 3D virtual se localizan los nodos espaciales de la distribución de probabilidad de cada parte del cuerpo, y así se generan varias propuestas para las ubicaciones 3D de cada una de las articulaciones del cuerpo con cierto puntaje de confianza. La segmentación en partes del cuerpo se realiza con una clasificación por pixel para evitar la razón combinatoria que implicaría una búsqueda sobre las distintas propuestas de posiciones de articulaciones del cuerpo; esta debilidad del método se balancea con la gran cantidad de imágenes de profundidad de entrenamiento utilizadas.

Para lograr ello, los autores del SDK en una primera etapa crearon una base de datos de imágenes de profundidad reales de personas realizando distintos movimientos en condiciones ambientales diferentes, en las cuales las partes del cuerpo se etiquetaron manualmente. En base a ello generaron un modelo de movimiento humano en 3 dimensiones mediante el cual produjeron una gran cantidad de datos sintéticos, con imágenes de profundidad sintéticas de humanos de distintas formas y tamaños en poses muy variadas. Entrenaron un conjunto de árboles de decisión aleatorizados (Random Forest) evitando el sobre-entrenamiento y obteniendo invariancia a la traslación 3D gracias a la gigantesca cantidad de datos. Finalmente, utilizando el algoritmo de Mean-Shift se infieren los nodos espaciales de las distribuciones por pixel de las cuales se extraen distintas propuestas de las posiciones 3D de las articulaciones.

En todo el proceso no se emplea información temporal, ni datos extraídos de imágenes anteriores, lo cual lo vuelve robusto y permite una re-inicialización rápida cuando una persona sale y entra al campo de visión de la cámara.

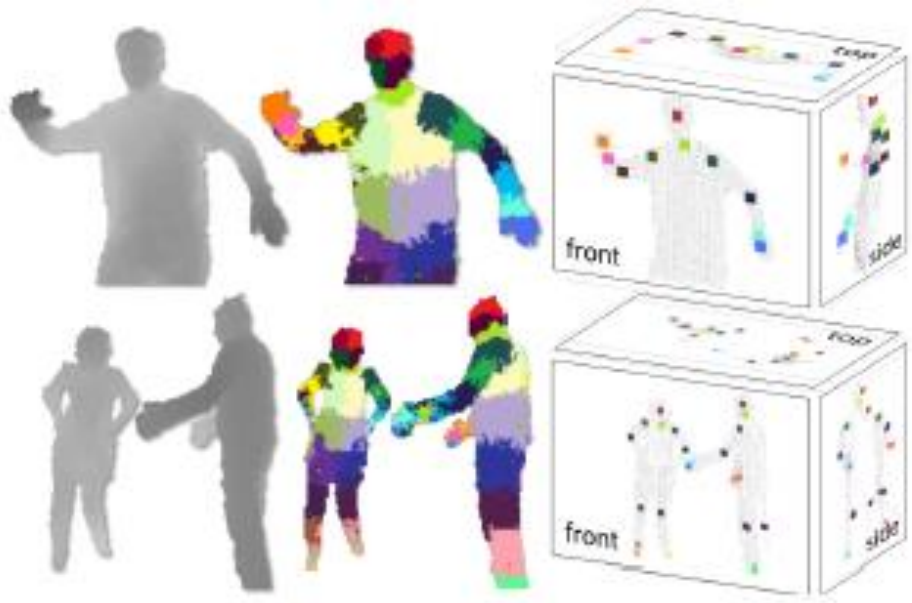


Figura 2.11 Imagen de profundidad, partes del cuerpo, modelos 3D propuestos. Fuente: [31]

2.5.7 Acerca del Sensor Kinect 2.0

Xataka [32] comenta que el dispositivo de captura llamado Kinect se ha convertido en mucho más que un método de control para consola de videojuegos, volviéndose un elemento importante de estrategia y referencia para distintos productos con variadas e interesantes aplicaciones.

Tres años después de la publicación de la primera reléase de la Kinect, Microsoft publicó la Kinect 2.0 la cual supuso una mejora respecto a su predecesor.

La diferencia principal reside en la incorporación de una nueva cámara principal, contando con una cámara Time-of-Flight (TOF) de alta resolución pudiendo capturar mayor detalle con precisión y mejor resolución. El nuevo modo de profundidad proporcionado por la cámara TOF permite reproducir una escena con tres veces más fidelidad que la Kinect 1.0.

Logra un campo de visión 60% más grande, registrando mayor espacio. Es capaz de reconocer hasta 6 personas en escena, en comparación con sólo 2 de su antecesor. Cuenta con un sensor de infrarrojos más potente, reconociendo objetos y personas en condiciones de muy poca luz o totalmente a oscuras. Detecta posturas de la mano hasta 4 metros de distancia, distinguiendo con precisión los dedos. El reconocimiento facial se ve ampliamente mejorado, detectando gestos de forma precisa.

Hasta 2 gigabits de datos por segundo son recogidos por el dispositivo al leer el entorno, requiriendo un software capaz de procesarlo y permitir la interpretación de lo escaneado, es aquí donde el apoyo de Microsoft Research jugó vital importancia en apoyo al equipo Xbox.

2.5.7.1 Time-of-Flight en Kinect 2.0

Xataka [32] explica que las cámaras con tecnología TOF emiten señales lumínicas que rebotan en los objetos y son recogidas de vuelta midiendo el tiempo que tardan en recorrer la distancia. Para que lo mencionado funcione adecuadamente distinguiendo reflejos de objetos en una habitación y su ambiente, es necesario una precisión de hasta 1/10 mil millones de segundos. Siendo esta la única forma de proporcionar suficiente información para procesar las formas y contornos de los objetos.

Durante el desarrollo del Kinect 2.0 se presentaron diversos problemas y todos debían ser solucionados antes de la fecha de publicación del Xbox One (Finales del 2013), convirtiéndose en aliado el equipo de investigadores Microsoft Research quienes lograron optimizar algoritmos, parámetros, superar la tarea de procesar todos y cada uno de los 220 mil pixeles que recoge el dispositivo y sus sensores incorporados.

Otra gran misión era medir con precisión pequeños objetos en cualquier escenario y bajo todo tipo de luminosidad. Finalmente lograron diferenciar dedos de manos del contexto

y en general, detectar objetos de sólo 2.5 cm., en comparación a los 7.5 centímetros de su antecesor.

Figura 2.12 Imagen de profundidad obtenida del Kinect 2.0. Fuente: [32]

2.5.7.2 Comparación entre Kinect 1.0 y 2.0

A continuación, se presenta un cuadro comparativo entre los sensores Kinect 1.0 y su sucesor:

Característica	Kinect 1.0	Kinect 2.0
Cámara de Color	640 x 480 / 30 fps	1920 x 1080 / 30 fps
Cámara de Profundidad	320 x 240	512 x 424
Máxima Distancia de Profundidad	~4.5 metros	8 metros
Mínima Distancia de Profundidad	40 centímetros	50 centímetros
Ángulo de Visión Horizontal	57 grados	70 grados
Ángulo de Visión Vertical	43 grados	60 grados

Motor de Inclinación	Sí	No
Cantidad de Joints Detectables	20 joints	25 joints
Cantidad de Esqueletos Detectables	2	6
Estándar USB	2.0	3.0
Sistema Operativo Soportado Oficialmente	Windows 7 y 8	Windows 8
Precio	50 dólares	250 dólares

Tabla 2.3 Comparación entre sensores Kinect. Fuente: [33]

Se ha demostrado la superioridad en diversos aspectos del sensor Kinect 2.0 en comparación a su antecesor, resaltando las ventajas de proveer mayor detalle en el reconocimiento de partes minúsculas como los dedos y seguimiento del rostro, siendo estos de vital importancia en el desarrollo y diferenciación de la mayoría de gestos, por lo cual se optaría por considerar al sensor Kinect 2.0 como ideal para la implementación de un software intérprete de lengua de señas. No obstante, considerando factores determinantes como el alcance definido para este trabajo, siendo por el momento innecesario el detalle a precisión de partes minúsculas del cuerpo humano, el factor económico, la compatibilidad con sistemas operativos mayormente usados, la necesidad de puerto USB 3.0 y la facilidad de adquisición dentro de los plazos establecidos, se decide trabajar con el Kinect 1.0.

2.6 Acerca de las Técnicas de Reconocimiento

2.6.1 Dynamic Time Warping

En Müller [34], se comenta que DTW es una técnica muy conocida para homologar o alinear óptimamente dos series temporales de datos variantes en tiempo y/o velocidad. Empleado con frecuencia en aplicaciones de reconocimiento de voz para lidiar con el problema de velocidad del habla.

También se ha utilizado en aplicaciones como reconocimiento de gestos, minería de datos, visión y animación por computadora, vigilancia, alineación de secuencias de proteínas, ingeniería química, música, en áreas como medicina, ingeniería en general, entre otros.

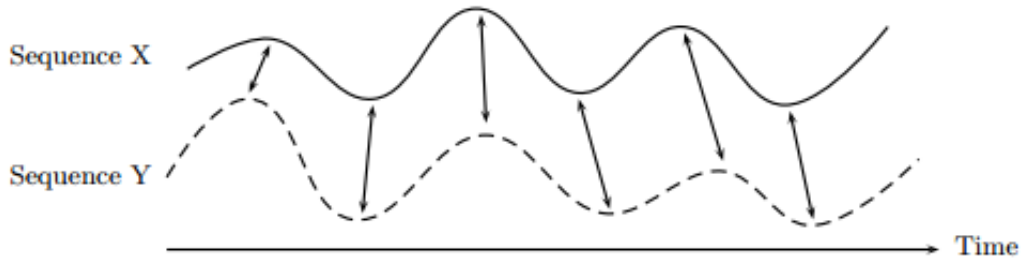


Figura 2.13 Alineación de tiempo de dos secuencias dependientes del tiempo. Fuente: [34]

Su objetivo es comparar dos secuencias $X = (x_1, x_2, \dots, x_N)$ de tamaño $N \in \mathbb{N}$ y $Y = (y_1, y_2, \dots, y_M)$ de tamaño $M \in \mathbb{N}$. Se define un espacio F donde $x_n, y_m \in F$ para $n \in [1{:}N]$ y $m \in [1{:}M]$. Para equiparar dichas secuencias se define una medida de costo definida como Distancia D:

$$c{:}\, FxF \rightarrow R_{\geq 0}$$

El valor $c(x, y)$ es pequeño si x y y son similares entre sí, caso contrario $c(x, y)$ es de costo alto. La medición del costo total se realiza con la comparación de cada par de elementos de las secuencias X y Y, obteniendo una matriz de costos $C \in R^{NxM}$ definido por $C(n, m) = c(x_n, y_m)$, donde la meta es encontrar un alineamiento que contenga el menor costo entre ambas secuencias.

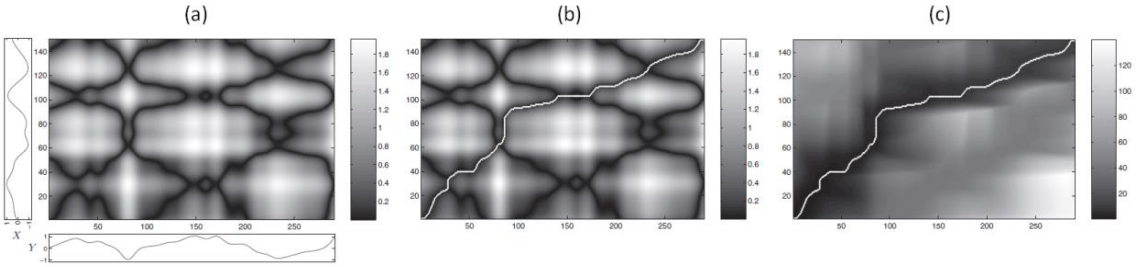


Figura 2.14 Matriz de costo (a) y Ruta de deformación óptima p^* (b y c). Fuente: [34]

La Matriz de Costo mostrada permite visualizar la comparación entre dos secuencias X y Y, donde las regiones de menor costo son las más oscuras.

Por lo tanto, el Alineamiento se define como una secuencia $p = (p_1, \dots, p_L)$ con $p_e = (n_e, m_e) \in [1{:}\,N]x[1{:}\,M]$ para $e \in [1{:}\,L]$ satisfaciendo las siguientes condiciones:

(i) Condición de Contorno: $p_1 = (1,1)$ y $p_L = (N, M)$.

(ii) Condición de Monotonicidad: $n_1 \leq n_2 \leq \cdots \leq n_L$ y $m_1 \leq m_2 \leq \cdots \leq m_L$.

(iii) Condición Tamaño de Paso: $p_{e+1} - p_e \in \{(1,0), (0,1), (1,1)\}$ para $e \in [1{:}\,L-1]$.

La Condición de Contorno obliga la comparación de los primeros elementos de las secuencias X y Y, al igual que los últimos.

La Condición de Monotonicidad busca prevenir que la comparación entre elementos de las secuencias X y Y retrocedan en el tiempo, esto quiere decir, si en el proceso de comparación se decide que un elemento x_i es comparado con y_j, entonces no es factible que algún punto de la primera secuencia con índice mayor a i se compare con un punto de la segunda secuencia con un índice menor a j, de forma similar, un elemento con índice menor a i de la primera secuencia no puede compararse con un punto mayor a j de la segunda.

Por último, la Condición Tamaño de Paso expresa un tipo de continuidad donde ningún elemento de X y Y debe ser omitido, ni repetido en el proceso de comparación si pudiesen ser equiparados, esto es el núcleo del método pero también el factor por el cual pierde su sentido óptimo.

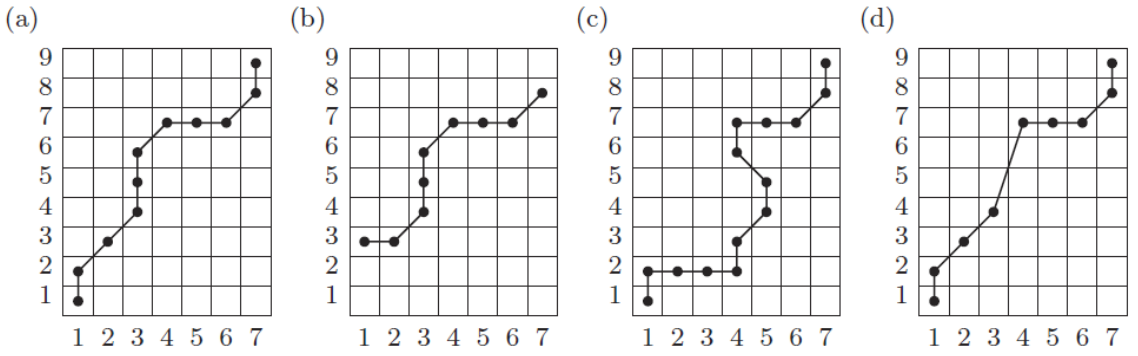


Figura 2.15 Rutas para secuencias X de tamaño 9 y Y de tamaño 7. Fuente: [34]

En la figura precedente se tienen cuatro casos que no necesariamente cumplen las condiciones para considerarse Rutas de Deformación. La imagen (a) cumple todas las condiciones. La segunda (b) incumple la Condición de Contorno al no empezar ni terminar en los límites. La tercera (c) viola el Condición de Monotonicidad. Por su parte, la última (d) no cumple con la Condición Tamaño de Paso.

Entonces, el costo total $c_p(X,Y)$ de una ruta de deformación p entre X y Y con respecto al costo local de medida c es definido como:

$$c_P(X,Y) \coloneqq \sum_{l=1}^{L} c(x_{n_l}, y_{m_l})$$

Donde $p*$ es la Ruta de Deformación Óptima conteniendo el menor costo total al comparar las secuencias X y Y:

$$DTW(X,Y) \coloneqq c_{p*}(X,Y)$$

$$= \min\{c_p(X,Y) \mid p\ es\ una\ ruta\ de\ deformación\ (N,M)\}$$

Para hallar $p*$ computacionalmente se requiere la evaluación de toda la matriz de costo D. La ruta óptima $p^* = (p_1, \dots, p_L)$ es hallada en orden inverso, es decir comenzando con $p_L = (N,M)$, se considera haber finalizado cuando se llega $(n,m) = (1,1)$, en caso contrario se sigue:

$$p_{l-1} \coloneqq \begin{cases} (1, m-1), & si\ n = 1 \\ (n-1, 1), & si\ m = 1 \\ argmin\{D(n-1, m-1), D(n-1, m), D(n, m-1)\} \end{cases}$$

En Choi [35] desarrollan una variante de DTW el cual según evaluaciones de los autores incrementa el porcentaje de reconocimiento al considerar no sólo la comparación de

valores de secuencia, sino también de posición y dirección. El avance de la tecnología y cambio de las interfaces de interacción tradicionales a comandos de voz y/o movimientos corporales a partir de sensores especiales, han generado mayor interés y desarrollo especialmente del HCI. En este aspecto, técnicas como Dynamic Time Warping (DTW) y Hidden Markov Model (HMM) han cobrado importancia como técnicas ideales para el reconocimiento de secuencias de datos, sin embargo, investigadores han visto conveniente mejorar el performance. DTW es capaz de realizar la comparación de dos secuencias de datos diferentes en tamaño, creando una matriz de costos por cada componente de las secuencias a partir del cual mediante programación dinámica se selecciona y guarda la de menor costo. En comparación a algoritmos basados en probabilidad, así por ejemplo HMM; DTW proporciona resultados constantes y más estables en comparación.

DTW es empleado en diversas áreas, tales como minería de datos, reconocimiento de movimientos o gestos y reconocimiento de voz. Se puede considerar como un algoritmo de coincidencia o comparación de patrones que permite una construcción no linear en escalas de tiempo definidas.

Así, Choi [35] menciona las siguientes bases matemáticas:

$$A = a_1, a_2, a_3, \dots, a_n$$

$$B = b_1, b_2, b_3, \dots, b_m$$

Además, se define D como la matriz formada por $(n\ x\ m)$ y la distancia de A y B obtenida al procesar:

$$d(w) = d(i,j) = ||a_i - b_i||$$

Las rutas de comparación de las dos secuencias se pueden expresar como:

$$W = w_1, w_2, w_3, \dots, w_k$$

Teniendo en consideración las rutas, se procede a buscar la de menor costo a partir de la siguiente ecuación, sin embargo, esta origina mayor costo computacional al comparar todas las secuencias posibles intentando encontrar la óptima.

$$DTW(A,B) = min \sum_{k=1}^{K} d(w_k)$$

Empleando la fórmula de recurrencia correspondiente a Programación Dinámica la cual rotacionalmente acumula la distancia mínima se permite un cálculo más rápido. Este puede expresarse como:

$$g(i,j) = d(i,j) + \min\left[g(i-1,j-1), g(i-1,j), g(i,j-1)\right]$$

2.6.2 Hidden Markov Models

Wang [36] define al HMM como un modelo estadístico, el cual incluye un conjunto finito de estados asociados a una distribución de probabilidad sobre todas las posibles salidas. Las transiciones entre estados dependen del conjunto de probabilidades asociadas. Dichos estados no son visibles, sin embargo, las salidas producidas sí.

Formalmente se define HMM por $\gamma = \{S, A, B, \pi\}$, y también mediante los siguientes parámetros:

- Un conjunto de estados $\mathrm{S} = \{1,2, \dots, \mathrm{K}\}$.
- Probabilidades de transición de estados $A = \{a_{ij}\}, 1 \leq i,j \leq K$, donde a_{ij} es la probabilidad de transición del estado i a j.
- Probabilidades de Salida $B = \{b_i(o)\}, 1 \leq i \leq K$. Donde o es una observación con un valor continuo o discreto, y $b_i(o)$ es la probabilidad del estado i generando la observación o.
- Probabilidades iniciales $\pi = \{\pi_i\}. 1 \leq i \leq K$. Donde π_i es la probabilidad de que las series tiempo comiencen en el estado i.

Un problema fundamental asociado al empleo de HMM se da al intentar encontrar la secuencia de estados óptima que produce las observaciones dado un modelo γ y una secuencia de observaciones O. Otro problema se da en la etapa de entrenamiento, al buscar estimar el parámetro γ, de tal forma que la probabilidad de la secuencia de observación generada por el óptimo estado de secuencia sea maximizada. Como solución se tiene que dado un HMM γ y una secuencia de observaciones O, la Producción de Probabilidad es la probabilidad de HMM γ generando O durante un estado de secuencias $s = (s_1, \dots, s_m)$ bajo la siguiente ecuación:

$$P(O, s|\gamma) = \pi_{s_1} b_{s_1}(o_1) \prod_{j=2}^{m} a_{s_{j-1},s_j} b_{s_j}(o_j)$$

Según Jiang [37], conocidos como Modelos Ocultos de Markov, es un proceso de Markov que se divide en dos componentes: Uno observable y otro no (Componente oculto). Un Modelo Oculto de Markov es un proceso de Markov $(X_k, Y_k)_{k\geq 0}$ en el espacio ExF, donde se asume que contamos con medios para observar Y_k, pero no X_k como la señal de proceso y E es la señal de espacio de estado, mientras el componente observado Y_k es llamado el proceso de observación y F es la señal de espacio de estado.

El objetivo es determinar los parámetros desconocidos de dicha cadena empleando los observables. Los parámetros extraídos se pueden emplear para llevar a cabo sucesivos análisis.

Ha sido ampliamente usado en reconocimiento de patrones, especialmente aplicados en el reconocimiento del habla, escritura manual, gestos, bioinformática, entre otros.

Los estados son representados usualmente por óvalos, donde cada uno es una variable aleatoria que puede tomar determinados valores. La variable aleatoria x(t) es el valor de la variable oculta en el instante t y sólo depende de x(t-1) en el instante t-1, la variable aleatoria y(t) es el valor de la variable observada en el mismo instante de tiempo y depende sólo de x(t), mientras que las flechas indican dependencias condicionales.

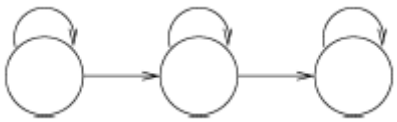


Figura 2.16 Estructura básica de HMM. Fuente: [37]

Un Modelo Oculto de Markov se define formalmente como una tupla (Q, V, π, A, B):

- El conjunto de estados $Q = \{1,2, \dots, N\}$. El estado inicial se denota como q_t. Cada valor t hace referencia a una posición.
- El conjunto V de posibles valores $\{v_1, v_2, \dots, v_M\}$ observables en cada estado M es el número de alternativas posibles y cada una hace referencia a una diferente.
- Las probabilidades iniciales $\pi = \{\pi_i\}$, donde π_i es la probabilidad de que el primer estado sea Q_i.
- El conjunto de probabilidades $A = \{a_{ij}\}$ de transiciones entre estados.
    - $a_{ij} = P(q_t = j | q_{t-1} = i)$, es decir, a_{ij} es la probabilidad de estar en el estado *j* en el instante *t* si en el instante anterior *t-1* se estaba en el estado *i.*

- El conjunto de probabilidades $B = \{b_j(V_k)\}$ de las observaciones.
    - $b_j(V_k) = P(O_t = V_k | q_t = j)$, es decir, la probabilidad de observar V_k cuando se está en el estado *j* en el instante *t*.

Las secuencias observables se denotan como $O = (o_1, o_2, \dots, o_T)$.

CAPÍTULO III: ESTADO DEL ARTE

Según Capilla [31], actualmente existe un gran número de Lenguas de Señas en el mundo (Más de 50), casi todos los idiomas oficiales hablados cuentan con una, así por ejemplo existe Lengua de Señas Americana (ASL), Lengua de Señas Mexicana (LSM), Lengua de Señas Francesa (LSF), Lengua de Señas Italiana (LIS), Lengua de Señas Irlandesa (IRSL), Lengua de Señas Británica (BSL), Lengua de Señas Australiana (Auslan), Lengua de Señas Alemana (DGS), Lengua de Señas Israelí (ISL) y Lengua de Señas Española (LSE), entre otros. A pesar de esta larga lista, la Lengua de Señas Americana es la más estudiada y cuya gramática ha sido aplicada con éxito en otras.

Acorde a Yin [4], en las últimas décadas el reconocimiento automático de gestos ha sido tarea de investigación activa en el área de Visión por Computador[13]. Varios proyectos han aprovechado el potencial de procesamiento a bajo costo ofrecido por las computadoras actuales para obtener mejores resultados. Dichas investigaciones en su mayoría han sido impulsadas por empresas comerciales dominantes o laboratorios de investigación de las más importantes universidades alrededor del mundo.

Según Bulbul [38], el creciente número de aplicaciones ha sido proporcional al número de proyectos multidisciplinarios en áreas como Interacción Humano Computadora (HCI), cuidado de la salud, entrenamiento deportivo, entre otros.

Yin [4] comenta que los proyectos de Reconocimiento de Gestos han empleado herramientas de diferentes naturalezas. Entre estos resaltan los guantes con colores específicos provistos de acelerómetros y otros sensores con la intención de facilitar la adquisición y procesamiento de datos.

La extracción y clasificación de características obtenidas de imágenes se centra en el reconocimiento del color, forma y movimiento. En investigaciones iniciales usualmente se recurría al color de guantes como facilitador para el reconocimiento y segmentación

[13] Disciplina científica, también conocida como Visión Artificial, incluye métodos para adquirir, procesar, analizar y comprender las imágenes del mundo real con el fin de producir información numérica o simbólica para que sea tratada por computadores.

de las manos del usuario, requiriendo en algunos casos el empleo de prendas con mangas largas.

Sin embargo, según Jiang [39] los estudios que han empleado imágenes RGB convencionales no han obtenido resultados robustos, puesto que la sensibilidad óptica se ve afectada principalmente por los niveles de luminosidad y fondos desordenados. En cambio, la data obtenida por una cámara de profundidad genera resultados más estables ante ruidos y cambios en iluminación. La información de profundidad obtenida resulta útil para calcular la distancia entre las manos y el cuerpo ubicados ortogonalmente en el plano de imagen permitiendo distinguir gestos ambiguos. Además, en contraposición al elevado costo de los sensores de profundidad directos, han surgido otros con similares capacidades, pero económicamente accesibles. En este aspecto resalta la Microsoft Kinect, un sensor de profundidad ampliamente empleado en la identificación de gestos cuyo software provee el Skeleton Tracking[14] para un mejor reconocimiento de expresiones humanas en comparación con usar sólo data de profundidad. No obstante, la eficacia del método empleado en el reconocimiento de gestos no depende sólo de las características adquiridas de imágenes, sino también de los Clasificadores de Gestos que se empleen.

3.1 Taxonomía

La ACM en [59] explica que, con la intención de categorizar los diversos trabajos de investigación en la disciplina Informática, un grupo de colaboradores de dicha entidad propusieron la Computing Classification System (CCS) como sistema de clasificación estándar y cuya última revisión fue publicada en el 2012. CCS es dinámico y se va actualizando al reflejar los distintos conceptos y categorías existentes y nuevas en el estado del arte.

Empleando la Clasificación de ACM, la solución propuesta en este trabajo se puede enmarcar según el siguiente diagrama:

[14] Permite a la Kinect reconocer personas y seguir sus acciones estando parados o sentados. Requiere que el usuario esté frente al sensor y el rostro sea detectado. El software provee la posibilidad de representar la información obtenida mediante un array de joints donde son representadas las principales articulaciones del cuerpo reconocido de tal forma que interconectadas se asemejan a un esqueleto humano.

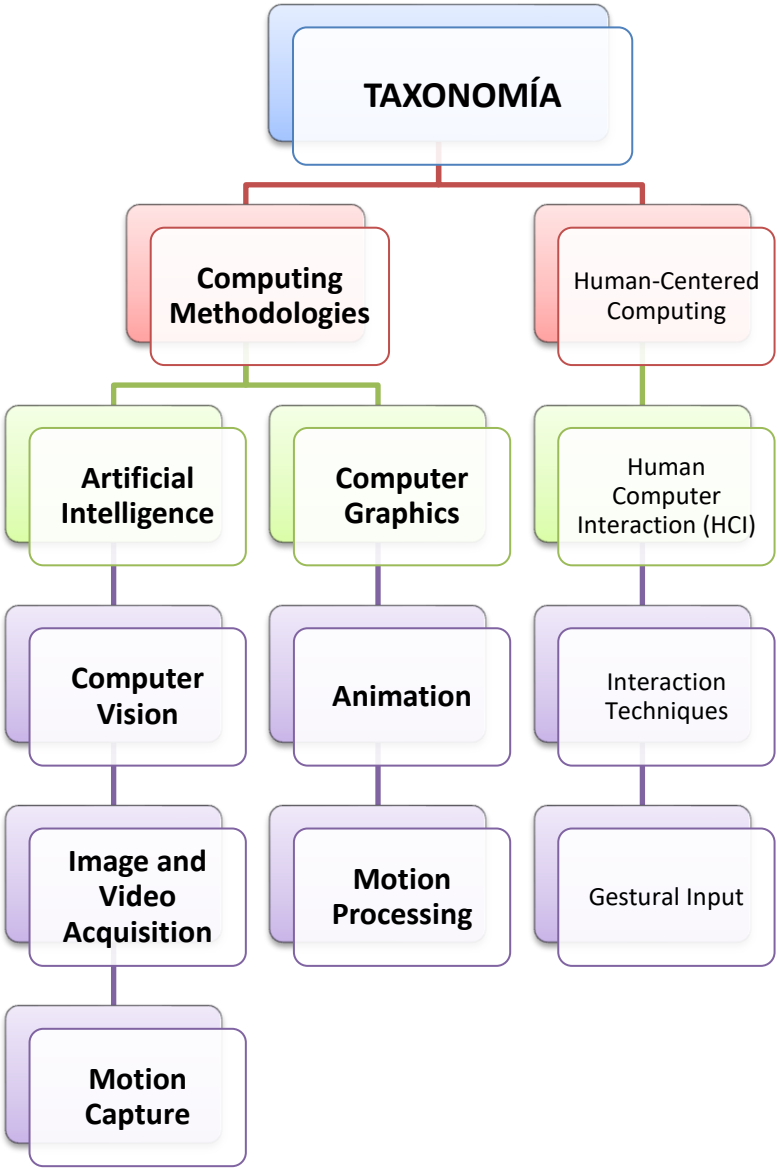


Figura 3.1 Clasificación de la solución propuesta según ACM. Fuente: Elaboración propia.

El texto resaltado en negrita denota mayor relevancia de las materias en la clasificación taxonómica.

Según la clasificación de ACM en [60], se ha visto por conveniente enmarcar el trabajo bajo dos grandes ramas:

a) Metodologías de la Computación (Computing Methodologies). Implica conocimientos en procesamiento de imágenes, visión por computador y comprensión de la escena, aplicaciones de Inteligencia Artificial y sistemas expertos.

b) Computación Centrada en el Ser Humano (Human-Centered Computing). El número de investigaciones en dicho campo ha crecido notablemente en los últimos años. Abarca conocimientos relacionados con interfaces y usabilidad.

3.2 Adquisición de Datos y Extracción de Características

La adquisición de datos es el primer paso crucial en un Sistema de Reconocimiento de Gestos. Si se incurre en una clasificación, el significado de los signos varía y deriva principalmente de las características Manuales, es decir la ubicación, movimientos, orientación y postura de la mano en un momento dado. Mientras que las características No Manuales se relacionan con las expresiones faciales, movimientos de los labios y posturas del torso. Esto último no está dentro del alcance de la presente investigación.

Según Galvez [40], el capturar y realizar seguimiento al movimiento del ser humano de manera computacional es una tarea siempre compleja con alternativas de solución costosas. Frente a ello, en las últimas décadas ciertos grupos de investigación de Reconocimiento de Gestos han empleado imágenes 2D captadas por cámaras convencionales, otros han propuesto el uso de datos 3D obtenidos por cámaras de profundidad, esta tendencia ha ido creciendo conforme se han abaratado los costos de dichos instrumentos.

En esta tesis se plantea desarrollar un sistema que permita reconocer gestos a partir de datos 3D obtenidos por el sensor Kinect, de tal forma quede evidenciada su eficiencia en contraste a trabajos realizados mediante cámaras convencionales 2D.

3.2.1 Métodos de Seguimiento

Los gestos no están diseñados, ni condicionados para ser reconocidos automáticamente, estos pueden contener movimientos rápidos y ser ocluidos durante su realización. Gracias a estas dificultades los investigadores han propuesto diferentes métodos como alternativas de solución.

Isikligil [41] comenta que algunos de los métodos preferidos para la adquisición de datos es el empleo de sensores wearable y accesorios. En esta categoría se encuentran los guantes especiales de datos y acelerómetros. Por ejemplo, Waldron y Kim en “Isolated ASL Sign Recognition System for Deaf Persons” desarrollaron un sistema cuyo input es la información de seguimiento provista por un guante de datos con sensor Polhemus,

empleando además redes neuronales de dos capas. Vogler y Metaxas en "ASL Recognition Base don a Coupling Between HMMs and 3D Motion Analysis" usan sensores magnéticos con técnicas de visión por computador para realizar el seguimiento de las muñecas con precisión. Mientras Hernandez-Rebollar en "A Multiclass Pattern Recognition System for Practical Finger Spelling Translation" presentó un complejo sistema compuesto por guantes de datos y acelerómetros llamado Guante Accele, este sistema era capaz de unir posición y movimiento de los dedos y manos. Otro método empleado con frecuencia es el uso de guantes coloreados de forma especial por regiones. Este enfoque busca ayudar en el reconocimiento y segmentación de la imagen.

Basándose en que las principales características de un gesto son realizadas por la mano dominante, Zhang en "A Vision-Based Sign Language Recognition System Using Tied-Mixture Density HMM" usó un guante con diferentes colores para la mano discriminativa y un color particular para la otra. Los dedos, palma y parte trasera de la mano del guante dominante fueron indicados con 7 diferentes colores para facilitar una detallada extracción de características.

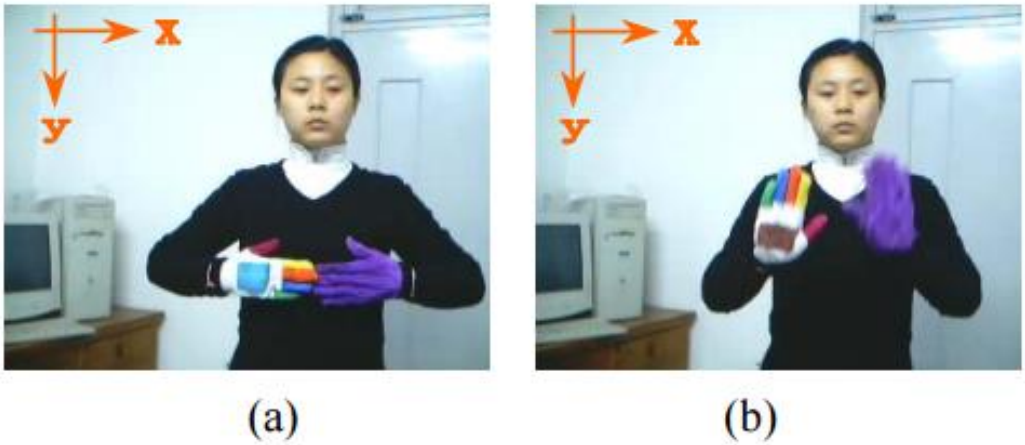


Figura 3.2 Vista frontal y posterior de los guantes de datos propuestos por Zhang. Fuente: [41]

Sin embargo, Isikligil [41] menciona además la extracción y segmentación basada en el color de la piel como la opción más natural para el seguimiento y reconocimiento de gestos, ello desde que algunos sistemas o propuestas requirieron el empleo de accesorios y sensores específicos volviéndose altamente intrusivos.

Como solución, mediante la publicación de herramientas como el sensor Kinect gran cantidad de investigaciones de reconocimiento de gestos se centraron en su empleo. Las ventajas de este artefacto al contar con una cámara de color integrada, un emisor de rayos

infrarrojos y un framework el cual permite realizar el seguimiento del esqueleto humano (Skeleton Tracking) ha desembocado en convertirla en una de las mejores opciones en adquisición de data para el reconocimiento de Lengua de Señas y gestos en general. Así, se tiene el trabajo realizado por Lang, Block y Rojas quienes en "Sign Language Recognition Using Kinect" crearon un sistema que emplea data y características otorgadas por el Skeleton Tracking para entrenar un HMM y posteriormente reconocer gestos. Mientras Doliotis en "Comparing Gesture Recognition Accuracy Using Color and Depth Information" usó la Kinect como cámara de profundidad para comparar el porcentaje de reconocimiento de gestos al emplear color y datos de profundidad realizando pruebas en un conjunto de gestos definido previamente.

Doliotis en [42] comenta que las aplicaciones de HCI y reconocimiento de gestos tienen como objetivo común proveer un enfoque natural de comunicación entre humanos y máquinas. Sin embargo, estas aplicaciones necesitan superar la complejidad de operar en ambientes naturales con fondos desordenados, objetos en movimiento y variaciones en condiciones de iluminación.

En la siguiente imagen, resultado de la investigación de Doliotis [42], se muestra el reconocimiento de manos basados en el color de piel. Se aprecian dificultades para detectar la mano cuando se hace un recorrido desde el rostro, además existen otros objetos en escena con color similar y los cuales son confundidos.

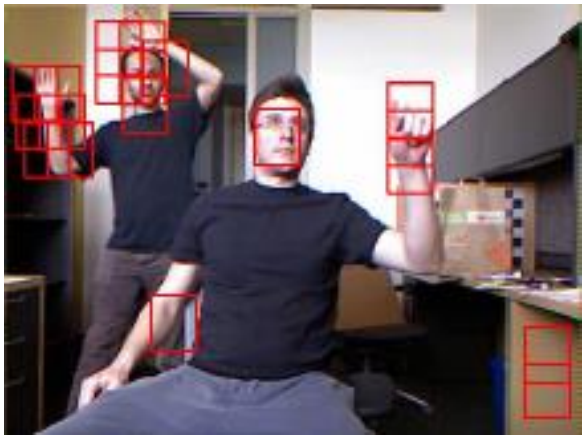


Figura 3.3 Detección de regiones candidatas según color de piel. Fuente: [42]

3.2.2 Evaluación

Considerando la problemática y objetivo del presente trabajo, así como la clasificación implícitamente dada a las herramientas mencionadas que han sido empleadas por investigadores como medio para captar datos, procesarlos y obtener resultados, a continuación, se explica el por qué se elige el empleo de una Cámara de Profundidad para el desarrollo del presente trabajo.

Galvez [40] resalta que los movimientos humanos son diversos y la cantidad de interpretaciones que se les puede asociar a los mismos es más extensa.

Acorde a Yang [27] y Bulbul [38], el reconocimiento de gestos humanos es una tarea compleja debido a factores como las dimensiones del cuerpo humano, posturas y seguimiento de los brazos, cercanía y movimientos de la cámara, ángulos de visión, condiciones de iluminación, entre otros. Además, la misma acción o gesto puede ser realizado de diferentes formas por diversas personas, incluso la misma persona diferentes veces.

Por otro lado, para Dong [10] los factores como la complejidad del gesto, oclusión de las manos durante el desarrollo del gesto y la limitada resolución del sensor empleado como input deben considerarse como influyentes en el resultado obtenido.

A lo largo de los años gran cantidad de investigadores han direccionado los problemas mencionados mediante el uso de características extraídas de imágenes 2D [43] [44], estas al ser capturadas por cámaras convencionales RGB no son capaces de brindar la información requerida y suficiente para realizar un análisis integral y comprensivo, además, se vería seriamente afectado por condiciones de iluminación. Se debe tener en cuenta que la identificación de los puntos principales pasa a depender de la textura del objeto en lugar de su geometría [38] [45]. Además, el empleo de imágenes 2D cuentan con limitaciones al ser necesaria la substracción del fondo y posterior segmentación del objeto en cuestión. [38]

Algunos autores plantean realizar el registro tridimensional de los movimientos empleando un mecanismo basado en múltiples cámaras 2D que trabajan paralelamente en el proceso de captura del movimiento permitiendo realizar un análisis desde diversos planos conllevando a una posterior integración de los mismos. Sin embargo, este enfoque recurre a herramientas y métodos que cuentan con un costo computacional y de desarrollo elevado que incrementa proporcionalmente al nivel de precisión deseado. Además, en la

actualidad con el avance de la electrónica e informática se pueden diseñar algoritmos potentes y procesarlos sin mayor esfuerzo computacional, denotando que las posibilidades de desarrollo han crecido exponencialmente, sin embargo, el registro del movimiento humano sigue siendo una tarea computacional compleja y costosa cuando se limita al ámbito bidimensional.

Recientemente, el surgimiento de las cámaras de profundidad a bajo costo han facilitado las tareas de seguimiento y reconocimiento significativamente, tal es el caso de la Microsoft Kinect, logrando menguar algunas de las dificultades mencionadas. [38]

Las imágenes producidas por una Cámara de Profundidad pasan a llamarse Imágenes de Profundidad, estas preservan la información de profundidad correspondiente a la distancia desde la superficie del objeto hasta el punto desde donde observa la cámara. Estas imágenes de profundidad permiten realizar análisis robustos en comparación con las imágenes de color debido a que factores como la iluminación, cambios de textura, fondos desordenados y otros no afectan considerablemente. Además, capturan la estructura 3D de la escena, también de la mano con su desplazamiento 3D, así como de los sujetos u objetos en cuestión. [38]

Valiéndonos de lo mencionado, considerando principalmente la necesidad de aprovechar el bajo costo computacional y económico, las cámaras de profundidad se muestran idóneas al contar con ventajas sobre las RGB convencionales. Entre dichas ventajas se pueden resaltar la capacidad de trabajar bajo condiciones mínimas de iluminación o en oscuridad total, otorgando un rendimiento constante ante variaciones de color y texturas que pueda mostrar el objeto en escena. También supera las ambigüedades en las tareas de visión por computador, como la sustracción de fondo y eficiente segmentación del objeto. [38]

3.3 Técnicas Propuestas por Terceros

En el punto 3.2 se mencionaron algunas técnicas empleadas a lo largo de los años por investigadores enfocados en el Reconocimiento de Gestos, donde se denota la importancia de la adquisición de datos y extracción de características para el proceso de reconocimiento de gestos. Sin embargo, esto no sólo depende de ello, sino también de la eficacia del Método de Clasificación empleado. Por lo tanto, de similar forma a lo visto en el punto anterior, a continuación, se presenta un resumen de los métodos de

clasificación y diferentes configuraciones que han sido propuestos en las investigaciones de Reconocimiento de Gestos. [41]

Según Yang [27], desde inicios del reconocimiento de gestos, los investigadores han empleado diferentes tipos de sensores para obtener información, es decir, diversos tipos de dispositivos de entrada, tales como: Cámaras de color, estéreo cámaras, guantes de datos, cámaras Time of Flight (TOF), etc.

Los sensores se pueden clasificar acorde a su ubicación para la obtención de información en el reconocimiento de gestos, así se tiene a los denominados Sensores de Entorno y los Sensores Wearable. Los primeros se caracterizan por estar instalados y capturar información a partir de una posición definida e independiente al usuario, tales como los sensores Kinect y Leap Motion Controller. Mientras que, los Sensores Wearable se ubican en las prendas de usuarios, aquí resaltan los sensores IMUs.

Para hacer frente a la problemática de conocer los ángulos formados por las articulaciones de los dedos de cada mano, los investigadores desarrollaron y emplearon técnicas mecánicas. En esta metodología resalta el enfoque basado en guantes, tales como CyberGloves y Powergloves [4]. Todos los tipos de guantes tomados como inputs han sido creados para determinar las posiciones de los dedos y palmas en el espacio 3D, por lo tanto, facilitar el seguimiento de la mano y reconocimiento de los gestos, este enfoque produjo buenos resultados en aplicaciones de reconocimiento de lengua de señas en décadas pasadas [46].

En Lee [47] se presentó un sistema de reconocimiento que consta de cuatro etapas, en la primera fase se obtiene la data a partir de la posición de los dedos y orientaciones de las manos mediante CyberGlove y sensores Polhemus. El sistema realiza la segmentación continua del gesto para extraer los signos e ignorar los movimientos que carecen de sentido y son realizados durante la finalización de un gesto y el inicio de otro. Luego, los movimientos de las manos se clasifican mediante la extracción de características y reglas de lógica difusa. Finalmente, se emplea un clasificador FMMNN para lograr reconocer 131 palabras y 31 letras del alfabeto manual de KSL. Los autores logran un reconocimiento del 80.1%.

Cabe resaltar que el empleo de dichos guantes ha logrado mejor desempeño y resultados que otras propuestas, sin embargo, el ser caros e intrusivos han limitado su popularidad. La combinación de imágenes de color y data de profundidad (RGB-D) ha sido empleada

exitosamente para demostrar el aumento de robustez en condiciones de luminosidad variable y reducir costos computacionales.

En los últimos 20 años, se han construido gran variedad de guantes sensores como input para aplicaciones de Interacción Humano Computadora, algunos de ellos se han mantenido en los laboratorios de desarrollo y otros han llegado al mercado [46]. Sin embargo, resultaron difíciles de emplear fuera de áreas de desarrollo debido a que proveen una experiencia no natural (Intrusiva) en la gesticulación, contando además con una configuración difícil y elevados costos [10], volviendo el procedimiento de detección más engorroso.

El abaratamiento de los dispositivos de interacción natural ha logrado impulsar la investigación y desarrollo de nuevas aplicaciones basadas en reconocimiento de gestos. [48]

Actualmente, el empleo de dispositivos de Interfaz Natural (NUI) para el control o interacción mediante gestos se ha vuelto común en nuestra vida cotidiana. En el 2006, el lanzamiento de la Wiimote cambió el concepto de control remoto, facilitando detectar movimientos de la mano en un espacio 3D. [48]

Gracias a ello se planteó otro enfoque para el reconocimiento de gestos, el cual consistía en incorporar información de distancia a los objetos o profundidad, empleando normalmente cámaras 3D o un conjunto de cámaras que generen imágenes 3D. [46]

Por otra parte, un enfoque que implica guantes y cámaras RGB para la detección ve forzada la coloración del guante para distinguir eficientemente las partes de la mano, reduciendo su aceptación por parte del usuario. [4]

La forma correcta e idónea de realizar el reconocimiento de gestos es con manos descubiertas, proveyendo experiencias naturales, sin embargo, el nivel de reconocimiento de gestos puede verse limitado. [10]

Las principales limitaciones para el reconocimiento mediante empleo de cámaras RGB son productos de factores externos, como consecuencia los investigadores han buscado diferentes tipos de inputs que puedan ser empleados para dicho propósito. [4]

Investigaciones como Oka [49], emplearon imágenes térmicas para realizar la segmentación de la mano con el fondo y la variación de iluminación, bajo la premisa de que la temperatura de la mano es casi siempre distinta a la del contexto.

Artículos como Li [46] afirman a través de investigaciones en trabajos de terceros que el reconocimiento de gestos realizado con videos 2D y algoritmos de seguimiento de manos empuja a confusión al ocurrir oclusiones entre diferentes partes del cuerpo, como consecuencia, la segmentación basada en color de manos no resulta fácil, viéndose influenciado por las condiciones de luminosidad y contexto. Entonces, el reconocimiento mediante guantes resulta capaz de proveer información de posiciones que puede ser un gran camino para mejorar la segmentación basada en colores, pero tal como ya se mencionó anteriormente, esta tiende a ser una experiencia inconfortable para el usuario. Es así que, un nuevo enfoque menos intrusivo ha venido tomando forma desde aproximadamente diez años empleando discontinuidades de profundidad[15] para separar la mano del fondo, de este modo los sistemas basados en profundidad evitan los problemas mencionados anteriormente.

La aparición y comercialización de aparatos como la Kinect y Leap Motion Controller han abierto las posibilidades del reconocimiento de gestos en el espacio 3D. La primera generación del sensor Kinect puede proporcionar un flujo de video de 640 x 480 píxeles a una frecuencia de 30Hz, o 1280 x 960 píxeles a 12Hz. El error de la medida de profundidad incrementa cuando se aumenta la distancia hacia el sensor. La Kinect puede realizar la detección y seguimiento de 20 puntos del cuerpo (Articulaciones o Joints), incluyendo las manos. Mientras que la segunda versión emplea la tecnología Time-of-Fligth, ofreciendo una resolución de 1920 x 1080 píxeles a 30Hz, siendo capaz de seguir hasta 25 puntos del cuerpo, incluyendo dedos. [4]

La aparición de la Kinect ha vuelto las tareas de reconocimiento de lengua de señas en general potencialmente factibles, acaparando gran interés de los investigadores en los últimos años. [10]

Mientras, Leap Motion Controller resulta ideal para la detección de movimientos de dedos (Cuenta con dos cámaras IR monocromáticas y tres LEDs infrarrojos), facilitando el seguimiento de los diez dedos de las manos. Sin embargo, las tareas de detección de movimientos de brazos y posturas de manos (Tales como: Mostrar la palma) pueden resultar un verdadero reto. [4]

---

[15] Diferencia entre datos de profundidad obtenidos.

Recientemente una nueva generación de Smart Watches está liderando el mercado de interfaces computacionales wereable. Estos dispositivos están equipados con sensores de movimientos tales como acelerómetros[16], giroscopios[17] y magnómetros[18] que son básicamente componentes de Unidad de Medición Inercial[19] (IMU). Estos sensores dan información con mayor precisión y la información de orientación de las manos con alta frecuencia de imágenes, por ejemplo, el sensor Xsens MTw IMU tiene una frecuencia de 50Hz y puede ser empleado como una entrada para los gestos. Mientras la data obtenida de una cámara basada en sensores es propensa a sufrir oclusión y la calidad de eficiencia depende de la posición del usuario frente al sensor, en contraste, los sensores IMU son independientes de ellos. Adicionalmente, la data otorgada por el IMU puede ser usada con un procesamiento menos complejo en comparación con los datos basados en los sensores de cámara. La desventaja del IMU es que no pueden capturar información del desplazamiento de las manos. [4]

Por motivos ya mencionados, en el presente trabajo nos centraremos en el empleo del sensor de Entorno llamado Kinect.

[16] Es un sensor de movimiento. Esta función la incorporan actualmente todos los teléfonos de última generación por la cual al mover el terminal la imagen de pantalla también se mueve.

[17] Empleado para medir o mantener la orientación. Basado en el principio de conservación del momento angular. Ha sido aplicado para disminuir el balanceo de navíos, aeronaves o proyectiles, estabiliza plataformas de tiro, suspensión de helicópteros, entre otros.

[18] Permiten cuantificar en fuerza o dirección la señal magnética de una muestra.

[19] O Intertial Mesurement Unit, es un dispositivo electrónico que mide e informa acerca de la velocidad, orientación y fuerzas gravitacionales de un aparato, usando combinación de acelerómetros y giroscopios.

3.3.1 Trabajos Previos

3.3.1.1 Sign Language Translator using Microsoft Kinect XBOX 360™

Capilla en [31] tuvo por objetivo desarrollar un traductor automático de lengua de señas el cual muestra la traducción (Palabra) del gesto ejecutado por un usuario ubicado frente la cámara (Input). La principal contribución de su trabajo es mostrar la eficiencia de Microsoft Kinect XBOX 360™ en combinación con un descriptor básico y un clasificador para la tarea de traducción de lengua de señas bajo un enfoque de ayuda social facilitando la interacción entre las personas que emplean lengua de señas para comunicarse y quienes son oyentes.

Para propiciar el entendimiento de su proyecto se muestra la siguiente imagen:

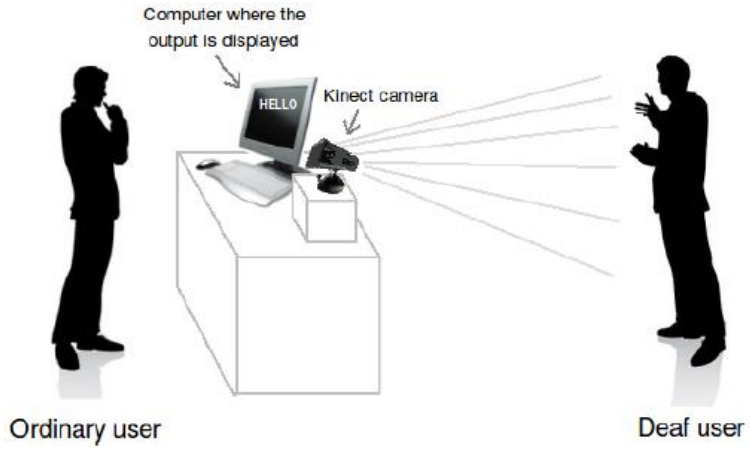


Figura 3.4 Objetivo del sistema propuesto por Capilla. Fuente: [31]

Trabajos previos emplearon modelos probabilísticos tales como Hidden Markov Model o Redes Neuronales Artificiales como clasificadores. Según el autor, dichos trabajos caían en complejas implementaciones basadas en descriptores estadísticos conllevando al aumento de costos computacionales. En contraposición, el autor propone el empleo de la Microsoft Kinect XBOX 360™ como alternativa frente a la problemática.

El software propuesto por el autor cumple con criterios de usabilidad, procesa el reconocimiento inmediatamente después de la ejecución del gesto y permite al usuario entrenar al sistema y agregar nuevos gestos.

El sistema comprende un Descriptor de 8 dimensiones para la data obtenida de cada gesto, adicionalmente compara resultados de eficiencia computacional entre los métodos de

clasificación Nearest Neighbor DTW y Nearest Group DTW, bajo distintas configuraciones.

Acerca de la metodología empleada, el autor propone la siguiente:

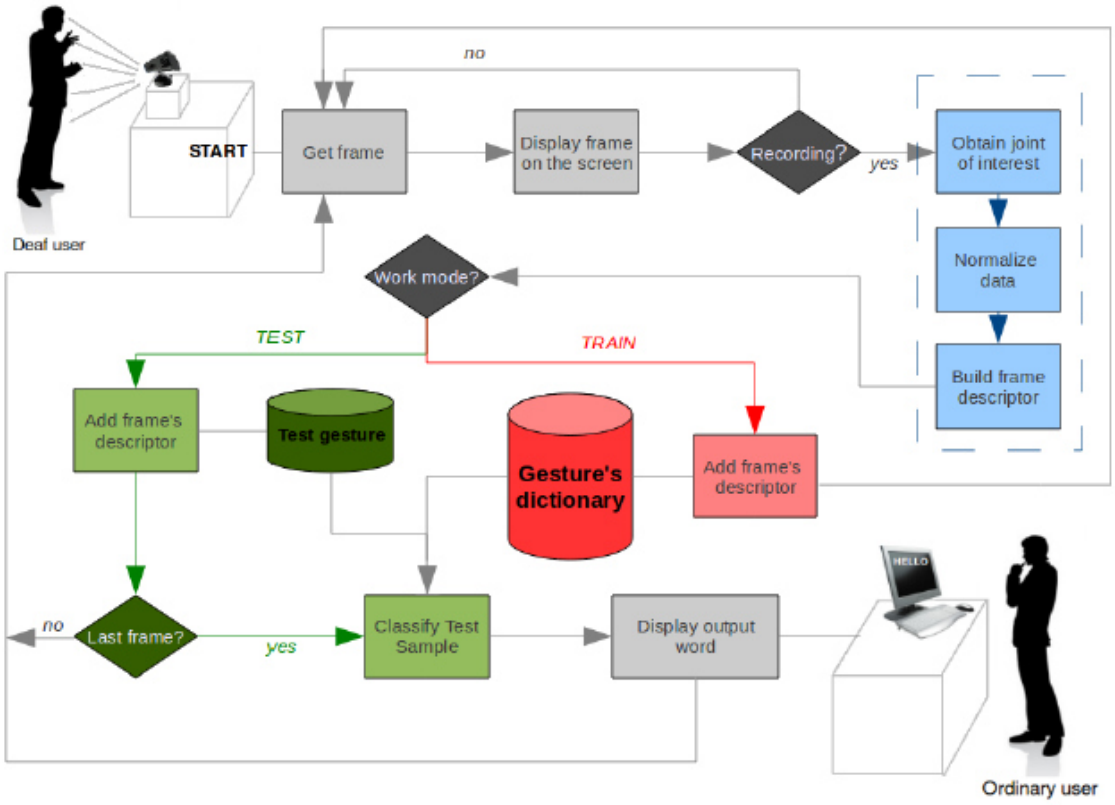


Figura 3.5 Diagrama de flujo del sistema propuesto por Capilla. Fuente: [31]

El usuario sordo (Deaf User) ubicado frente al sensor se encuentra realizando un signo o alistándose para ello, entonces se obtiene un nuevo frame (Imagen) y el flujo del video de entrada es actualizado con la imagen del esqueleto superpuesto encima del cuerpo humano en el video. Si el usuario desea Guardar, entonces se ejecutan de forma secuencial los tres bloques celestes, caso contrario el sistema consulta a la cámara para obtener el siguiente frame. El primer bloque celeste obtiene data de los joints (Articulaciones) de interés requeridos por el Descriptor, el segundo bloque tiene por función normalizar la información recibida y el tercero consiste en la construcción del descriptor de la trama. Entonces, si el modo de trabajo está definido como Training (El usuario desea agregar un nuevo gesto al diccionario) el descriptor es agregado al archivo del gesto correspondiente en el diccionario. De otra forma, si el sistema está en modo Testing (El usuario desea traducir el signo que se encuentra realizando), el frame es agregado al archivo "test.data". Posteriormente el sistema verifica si el actual frame es el último correspondiente al signo.

Luego que un signo es finalizado y si el modo de trabajo es Testing, entonces el signo realizado es comparado mediante el Clasificador con los signos almacenados en el diccionario y la respuesta es mostrada en lenguaje hablado del usuario ordinario. Finalmente, el sistema pasa al siguiente frame y el flujo se repite.

La normalización de la data de entrada tiene por finalidad robustecer el sistema ante diferencias de tamaño y la ubicación de quien realiza el gesto. Una ligera variación en esta última afectaría los valores del plano X y Y, gracias a las distancias entre los joints cambiarían considerablemente.

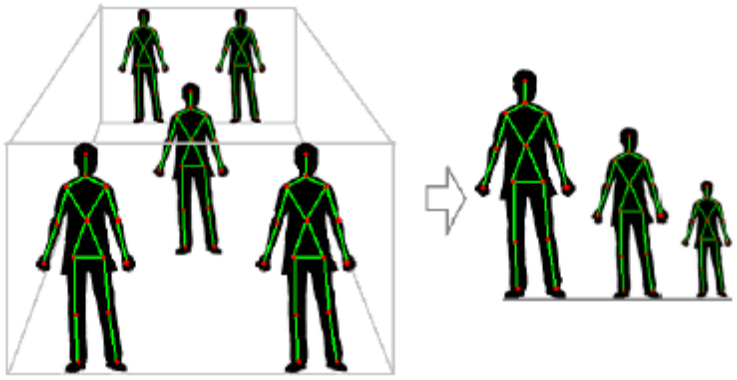
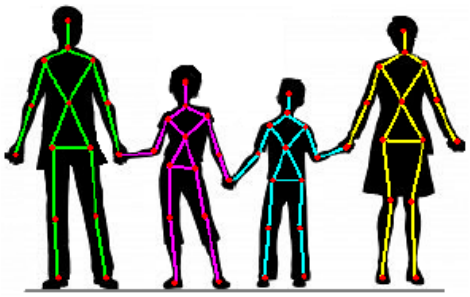


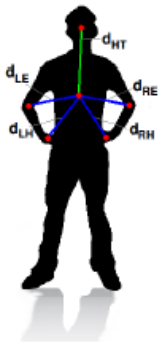
Figura 3.6 Normalización requerida por la posición del usuario propuesto por Capilla. Fuente: [31]

Frente a ello, el autor normaliza las coordenadas de los joint de interés con respecto a la ubicación tridimensional del Torso basándose en conceptos matemáticos del Sistema de Coordenadas Esférico. La posición del torso siempre es central con respecto al usuario.

Por otra parte, las diferencias entre las estaturas de los usuarios conllevan problemas al acarrear variaciones entre las distancias de joints. Como solución, cada uno de los joints son expresados por una relativa distancia desde el torso hasta ellos, lo cual considera los ángulos de orientación 3D. Además, se emplea un factor determinado por la distancia entre la cabeza y el torso.



(a) Different user sizes



(b) Set of distances D

Figura 3.7 Normalización requerida por el tamaño del usuario propuesto por Capilla. Fuente: [31]

Una vez normalizada la data se procede a construir un descriptor por cada signo, este debe ser único para cada signo y lo suficientemente diferente para ser distinguido de otros almacenados en el diccionario. El descriptor contiene tantas filas como frames comprenda el gesto, además se almacenan las coordenadas esféricas.

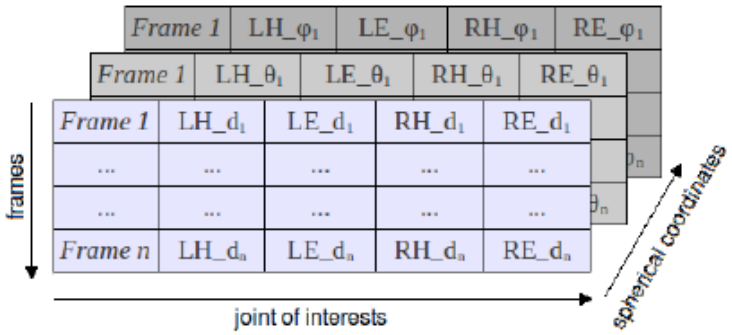


Figura 3.8 Descriptor de signos basado en Coordenadas Esféricas para cada joint propuesto por Capilla. Fuente: [31]

Por su parte, el Clasificador es el encargado de elegir el output acorde al signo realizado comparando cada signo con todos los almacenados en el diccionario. El autor maneja dos propuestas para el clasificador Nearest Neighbor DTW y el Nearest Group DTW. El primero, dado un signo en modo Test, el clasificador lo compara con el signo más cercano del diccionario con el fin de encontrar similitud entre el signo ingresado y cada uno de los definidos en el diccionario. La segunda propuesta es una versión modificada, diferenciándose en ser comparado con el más similar ejemplo del diccionario, donde el signo es contrastado contra el grupo más similar de ejemplos de signos del diccionario.

Cabe mencionar que el proyecto no se centra en un diccionario particular de signos porque su objetivo es evaluar la eficiencia de su propuesta, logrando una exactitud de

reconocimiento del 95.238% para el conjunto de signos definidos para sus pruebas, además de unas frases básicas compuestas por conjuntos de palabras establecidas.

Como trabajo futuro, el sistema empleará signos reales de lengua de señas, reconociendo la postura de los dedos y el movimiento de las manos. También propone implementar el reconocimiento automático del inicio y fin de cada gesto de forma continua. Se agrega el deseo de realizar mejoras a los algoritmos y costos computacionales acarreados al realizarse el proceso de reconocimiento de gestos.

Realizando un análisis de la propuesta de Capilla en [31], se tomará en cuenta el diseño de su interface al conversar con aspectos de usabilidad (Intuitivo, de fácil control y rápido aprendizaje), lo cual resulta clave si lo deseado es llevar el aplicativo a distintos ambientes. También se considera importante tomar el flujo de funcionamiento del software, el cual inicia al capturar los frames y termina con una respuesta entendible entregada al usuario, este servirá como marco de referencia sobre el cual se harán adaptaciones y mejoras para lograr los objetivos planteados. Cabe resaltar que la normalización de la ubicación y el tamaño del usuario también se tendrán en consideración. A diferencia de Capilla en [31], en este trabajo se plantea el empleo de palabras existentes en el diccionario de la Lengua de Señas Peruana para realizar las evaluaciones y validaciones.

3.3.1.2 Online Human Gesture Recognition from Motion Data Streams

Zhao [50] comenta que la reciente aparición de cámaras de profundidad a bajo costo han impulsado las investigaciones en el reconocimiento de movimientos del cuerpo. Esto ha facilitado el desarrollo de numerosas aplicaciones de visión por computador, de las cuales resaltan el entretenimiento electrónico, video vigilancia, monitoreo de pacientes, casas inteligentes, entre otros. Sin embargo, la motivación de los autores es proponer soluciones que superen resultados de terceros al resolver dos tareas claves en el reconocimiento de gestos. La primera, reconocer de forma continua gestos con flujos de data proveniente de movimientos no particionados o segmentados. Y el segundo consistente en diferenciar variaciones del mismo gesto buscando no incurrir en errores de reconocimiento.

Antes del desarrollo de la propuesta, los autores detallan tres conceptos (Gesto, Acción y Actividad) los cuales en la literatura no cuentan con definiciones claramente diferenciadas. Los Gestos son entendidos como los componentes atómicos y elementales empleados por el cuerpo humano capaces de transmitir algún significado, por ejemplo,

estirar o levantar los brazos. Las Acciones, son actividades compuestas por uno o varios gestos organizados temporalmente, como caminar y saludar. Las Actividades se refieren a las interacciones que pueden darse entre dos o más personas con o sin implicancia de objetos. Estos últimos no son considerados en el reconocimiento de gestos por emplear objetos, los cuales no son tomados en cuenta por la herramienta empleada (Kinect).

Para el primer problema, terceros realizan la segmentación del flujo de data antes del reconocimiento del gesto, sin embargo, ello cae en la necesidad de determinación del tamaño del segmento acorde a las características del flujo. El segundo problema consistente en identificar correctamente un gesto frente a variaciones que se pueden presentar en la realización, terceros han intentado resolverlo mediante segmentación del flujo de datos para su posterior comparación con patrones ya almacenados que tratan al gesto como unidad indivisible. Sin embargo, dichas propuestas cuentan con debilidades cuando existen variaciones no pudiéndose diferenciar correctamente, un segundo inconveniente resulta al intentar resolver lo anterior almacenando una plantilla por cada variación, generando ineficiencia en el reconocimiento de gestos para flujos de datos en tiempo real.

Mediante Structured Streaming Skeletons (SSS) o Esqueletos de Transmisión Estructurados, los autores resuelven los problemas mencionados subdividiéndolos en:

- Variación de antropometría y punto de vista o ubicación del ejecutor. Cada par de puntos (Articulaciones) es considerado como una parte del cuerpo humano, a partir del cual se realiza la normalización de sus distancias empleando otras características físicas del cuerpo del ejecutor, buscando reducir con ello posibles problemas de reconocimiento frente a variaciones causadas por datos de tamaño y ubicación de diferentes ejecutores.
- Variación de la tasa de ejecución. Para la solución de dicho problema, terceros han empleado mayormente la segmentación automática durante la extracción de características. Bajo el término “La Mejor Subsecuencia”, los autores definen como tal a la que más se parece al movimiento en ejecución según el último frame capturado. Agregado a lo anterior, mediante el empleo de SSS durante la ejecución de algún gesto se van calculando valores de distancia, estos son obtenidos de la diferencia entre la mejor subsecuencia y el último frame del movimiento en ejecución.

- Variación en el estilo del gesto ejecutado. Los autores hacen frente al inconveniente basándose en el uso de plantillas que almacenan los movimientos de las distintas partes del cuerpo humano a nivel unidimensional a diferencia de otros enfoques donde emplean las plantillas considerando al gesto como unidad. Este enfoque permite representar diferentes estilos de realización del gesto mediante variaciones en algunas partes unidimensionales del movimiento corporal sin caer en redundancia de datos.

La propuesta planteada por los autores consiste en dos partes: La primera, denominada Etapa de Entrenamiento (Offline Learning) cuya finalidad es la construcción del diccionario donde se almacenarán las plantillas de los gestos, y la Etapa de Predicción (Online Prediction) en la cual se realiza el reconocimiento del gesto en ejecución.

La Etapa de Entrenamiento se define en cuatro pasos:

1) Generación de Datos de Movimiento. Como se ha especificado anteriormente, los gestos pueden entenderse como la unión de varias trayectorias unidimensionales, donde cada una representa a una trayectoria específica recorrida por dos joints del Skeleton Tracking, considerando lo anterior, los datos son obtenidos de su posición y movimiento. Los valores a almacenar son generados de la distancia normalizada entre cada par de joints, logrando así convertirlos en invariantes ante la posición y antropometría del ejecutor.
2) Entrenamiento de Plantillas del Diccionario. Lleva por finalidad la creación de una base de datos en la cual estén almacenados las subsecuencias de gestos segmentados manualmente. Estos segmentos son posteriormente agrupados mediante un algoritmo para luego ser almacenados en el diccionario. En cada agrupación, el segmento con menor valor de distancia en comparación a otras es escogida como la plantilla. Cada plantilla o gesto es definido por un conjunto de movimientos unidimensionales almacenando la distancia entre un par de joints. Por ejemplo, en la siguiente imagen se observa una representación del deslizamiento o desplazamiento de una mano.

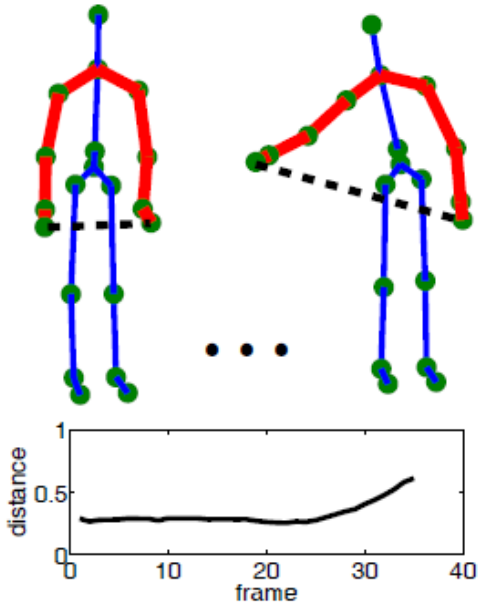


Figura 3.9 Representación de la normalización de distancias durante un tiempo establecido para el par de joints correspondientes a las manos. Fuente: [50]

Los autores consideran un acierto almacenar los movimientos de cada par de joints de forma individual para su posterior agrupamiento, como consecuencia con un pequeño número de agrupamiento pueden representar casi todos los movimientos del cuerpo humano. Si trabajaran con gestos de forma indivisible y considerando todas las partes del cuerpo, el número de agrupamientos se incrementaría reduciendo la eficiencia en el reconocimiento. Por lo cual, cada gesto es representado como una combinación de un número de plantillas de la base de datos, menguando redundancia de datos e incrementando la posibilidad de reutilizar lo existente.

3) Extracción de Características SSS. Una vez poblada la base de datos en los pasos anteriores, el conjunto de datos es escaneado con la finalidad de definir Vectores de Características SSS para codificar la información de los movimientos. Estos vectores son conjuntos de valores de distancia dados por el mínimo DTW identificado para cada par de joints (Unidimensional) que finalizan en el frame actual, emplear DTW permite eliminar la Variación de la Tasa de Ejecución. Así, si se cuenta con una plantilla acerca del movimiento de las dos manos, esta secuencia puede ser tomada en cuenta para realizar la comparación de todas las secuencias que incluyan dicho movimiento en esa dimensión. Este paso finaliza al convertir las etiquetas asignadas manualmente a las secuencias de frames en

vectores de características SSS asignados a cada frame, dejando listo al sistema para el entrenamiento de modelos de datos.

4) Entrenamiento de Modelo de Gestos. A partir de los vectores obtenidos se entrenan nuevos modelos de gestos, empleando para ello un clasificador llamado Jointly Sparse Coding.

Mientras la Etapa de Predicción se da en tres pasos, en primer lugar, los flujos de datos capturados a partir del Skeleton Tracking son convertidos en secuencias de datos de movimientos, posteriormente son transformados en vectores de características SSS, y finalmente, para cada frame es realizado el método de regresión lineal que asigna a cada vector de características una etiqueta según el modelo de gesto aprendido.

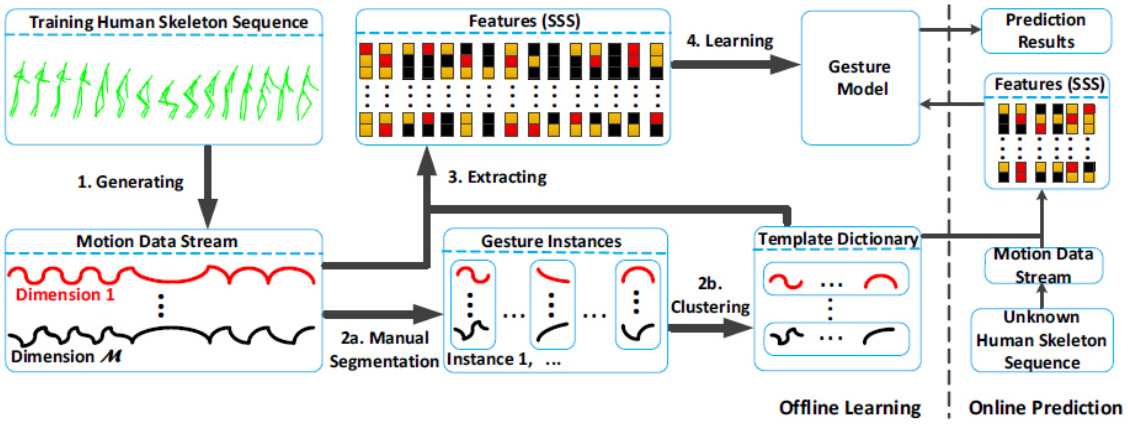


Figura 3.10 Framework o esquema de la propuesta de Zhao. Fuente: [50]

Para la evaluación de su propuesta, los autores plantean el empleo de un conjunto de datos de gestos denominados Kinect MSRC-12 y publicados en "Instructing People for Training Gestural Interactive Exterms" por Fothergill, Mentis y Kohli, conteniendo aproximadamente 700 mil frames correspondientes a 594 secuencias de 30 personas realizando 12 tipos de gestos. Antes de realizar la evaluación, los autores plantean el adiestramiento de los gesticuladores mediante las modalidades de: Descripción textual, secuencias de imágenes y videos de prueba. Finalmente, los autores denotan que las instrucciones de gestos brindadas mediante el empleo de sólo video tienden a producir peores resultados porque sus representaciones son consecuencia de sólo lo interpretado al observar los gestos.

Method	Text	Images	Video	Images+Text	Video+Text
Ours	$\mathbf{0.713 \pm 0.191}$	$\mathbf{0.666 \pm 0.194}$	0.557 ± 0.291	$\mathbf{0.730 \pm 0.148}$	$\mathbf{0.707 \pm 0.17}$
Baseline [8]	0.479 ± 0.104	0.549 ± 0.102	$\mathbf{0.627 \pm 0.053}$	0.563 ± 0.045	0.679 ± 0.035

Figura 3.11 Comparación del trabajo realizado versus el estado del arte de Zhang. Fuente: [50]

Adicionalmente, evalúan su enfoque de reconocimiento de gestos pre-segmentados empleando el conjunto de datos MSR-Action3D propuesto en "Action Recognition Based on a Bag of 3D Points" por Li, Zhang y Liu. Los autores emplean 20 clases de gestos mediante 10 personas logrando demostrar que su perspectiva de tratar gestos no como un conjunto completo e indiviso mejora la performance en comparación con otros métodos basados en series temporales.

Method	Accuracy
Recurrent Neural Network [19]	0.425
Dynamic Temporal Warping [20]	0.54
Hidden Markov Model [17]	0.63
Multiple Instance Learning [6]	0.657
Our Approach	**0.817**
Actionlet Ensemble [32]	0.882

Figura 3.12 Comparación de tasas de reconocimientos versus el estado del arte en Zhang. Fuente: [50]

El trabajo citado presenta distintas contribuciones, las cuales son:

1) Mediante Structured Streaming Skeletons (SSS) se logra el reconocimiento de gestos en tiempo real superando las limitaciones de antropometría, ubicación y estilos de realización del ejecutor.
2) A diferencia de otros, los autores no realizan segmentación previa al flujo de datos ingresante cuando un gesto se está ejecutando frente a la cámara. La propuesta implementada consiste en la detección automática del tamaño del segmento mediante una comparación dinámica con los patrones ya entrenados, reduciendo errores que puedan generarse gracias a segmentaciones previas.
3) Los autores definen y crean un diccionario o base de datos constituido por gestos a un nivel de granularidad compuesto por movimientos de cada parte del cuerpo. Ello se define así porque los gestos resultan de diferentes combinaciones del cuerpo humano.

4) Logran una buena performance de reconocimiento de gestos en tiempo real, llegando a obtener una latencia variante de 0.83 a 1.5 segundos para la cantidad de gestos empleados.

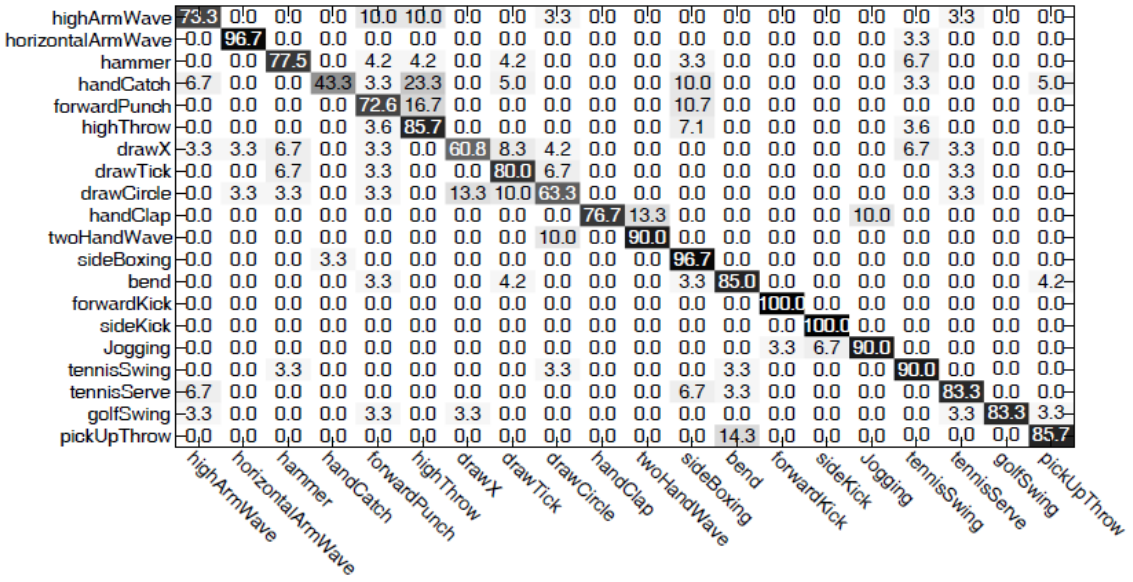


Figura 3.13 Resumen de porcentaje de reconocimiento según la propuesta de Zhang. Fuente: [50]

Como trabajo futuro platean el reconocimiento continuo de gestos en tiempo real empleando data inexacta del rastreo con la finalidad de reducir errores causados por data incompleta proveniente del Skeleton Tracking, así también el estudio sobre casos reales.

La propuesta de Zhao [50] comparte aspectos que resultarán de utilidad para el desarrollo del presente trabajo. Los registros (Gestos) almacenados en la base de datos o diccionario pueden ser muchos por cada tipo de gesto, sin embargo, es necesario definir a uno como plantilla que será empleado como referencia cada vez que se quiera clasificar un nuevo gesto. Para su definición Zhao [50] realiza una evaluación mediante el algoritmo DTW para todos los gestos entrenados por cada tipo de gesto, a partir del cual se elegirá al que muestre menor valor de distancia contra todos.

Zhao [50] va más allá de aspectos de software, considerando además para la validación de su propuesta el correcto adiestramiento de los gesticuladores, después de unas evaluaciones se determina que para un adecuado aprendizaje de los gestos a evaluar es necesario recurrir a videos o material didáctico que vaya de la mano con explicaciones orales de la ejecución. Además, para el entrenamiento del sistema y pruebas, los autores definieron y plantearon el uso de 20 tipos de gestos realizados por 10 personas, teniendo respuestas de procesamiento en tiempo real dentro de un tiempo que oscila entre 0.83 a

1.5 segundos. El desarrollo del presente trabajo ha de considerar los aspectos mencionados, es decir, para las evaluaciones, es necesario que quienes realizarán las gesticulaciones tengan el conocimiento adecuado sobre cómo ejecutar el gesto, en su defecto se recurrirá a videos y material didáctico para su aprendizaje. El tiempo de procesamiento del gesto también se tendrá en consideración, teniendo como umbrales los valores mencionados anteriormente.

3.3.1.3 Greek Sign Language Vocabulary Recognition Using Kinect

Gkigkelos [51] comienza mostrando la opinión de sus autores quienes aceptan el reconocimiento de lengua de señas como tarea desafiante en la traducción continua de gestos. Centran el trabajo en el reconocimiento de signos individuales mediante el empleo del sensor Kinect. La solución propuesta considera la evaluación de 15 signos de la Lengua de Señas Griega alcanzando una tasa de reconocimiento del 99.33%.

Aproximadamente el 5% de la población mundial, 360 millones de personas presentan pérdida de la audición, recurriendo al empleo de lengua de señas para transmitir ideas. Frente a esta realidad, varias entidades apoyadas con la tecnología, sensores y métodos algorítmicos han ideado y propuesto soluciones enfocadas en el reconocimiento de lengua de señas como forma de lidiar con las barreras de comunicación.

Aprovechando el multifacético empleo del sensor Kinect en áreas como Visión por Computador, Robótica y Reconocimiento de Gestos, los autores deciden emplearla en su propuesta esperando sea usada en contextos donde se requiera comunicación entre personas oyentes y no oyentes, sirviendo además de medio de entrenamiento para quienes desean aprender lengua de señas.

Terceros han empleado diversos métodos para la obtención de data proveniente de gestos, su posterior procesamiento y reconocimiento. Entre ellos se encuentran sensores wearables con bluetooth diseñados a medida y conveniencia de investigadores, pulseras negras cuya finalidad es servir de apoyo en la segmentación de imágenes, cámaras de escritorio básicas, cámaras montadas estratégicamente en un sombrero con la intención de registrar sólo el movimiento de manos, y el empleo de la Kinect. Las tasas de reconocimiento reportados oscilan entre 83 y 98.33%.

Con el objetivo de mejorar los métodos de terceros y obtener una mayor tasa de reconocimiento, los autores plantean el empleo del Kinect desarrollado para la Xbox One

y su respectivo SDK, el cual provee una mayor resolución de imagen y reconocimiento de joints en las manos.

La propuesta de los autores se enfoca en el reconocimiento de gestos individuales, para lo cual es necesaria una etapa de entrenamiento. El instrumento clave es el sensor Kinect, esta ha sido ampliamente empleada por otros investigadores al reconocer las facilidades de obtención de data de rastreo en tiempo real de partes del cuerpo.

El sistema se compone de diversos pasos iniciados al capturar frames que contengan data de un usuario realizando un gesto hasta que el usuario indique la finalización. Si el sistema se encuentra en modo Entrenamiento (Training), los frames se almacenarán bajo la etiqueta del gesto indicado, caso contrario en modo Test (Prueba) el signo se compara mediante un clasificador contra los existentes en la base de datos para arrojar un resultado acorde a mayor semejanza.

La segunda versión del sensor Kinect puede captar 25 joints de cada persona en escena, sin embargo, los autores reconocen la necesidad de sólo diez joints para el alcance planteado:

$$J = \{HTL, TL, WL, HL, EL, HTR, TR, WR, HR, ER\}$$

A lo anterior se añaden los joints de la cabeza (H) y torso (SM) considerados elementales en la normalización de la data obtenida.

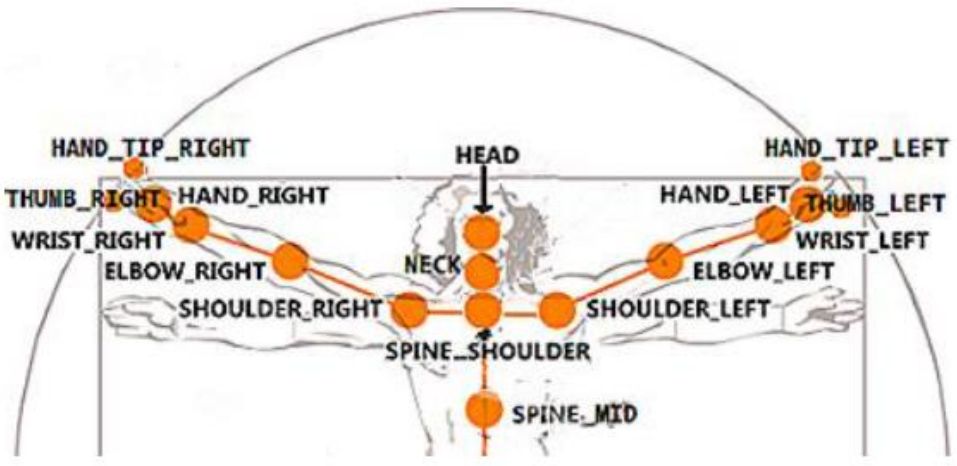


Figura 3.14 Joints seleccionados por el trabajo propuesto de Gkigkelos. Fuente: [51]

En búsqueda de mejores resultados, los autores consideran reducir el ruido presente en los frames correspondientes a los gestos realizados en modo Prueba y Entrenamiento, posteriormente realizan la normalización de la data para aumentar la tasa de

reconocimiento ante variaciones en las coordenadas X, Y y Z originadas por la ubicación del usuario escena, los cuales desembocan en diferentes valores para un mismo gesto. La normalización se lleva a cabo a partir del joint Spine Mid (SM) cuya posición es constante durante la realización del gesto. El empleo de los datos 3D es mediante el Sistema de Coordenadas Esféricas, este es un método muy empleado para representar figuras en tres dimensiones mediante tres coordenadas: La distancia radial de un punto hasta un origen fijo (R), el ángulo cenital desde el eje Z positivo (θ), y el ángulo Acimutal desde el eje X positivo $(\emptyset)$. Los valores tomados en consideración resultan de la distancia (r) entre los joints (J) y el centro origen, así se consideran los ángulos polares (θ) y azimutales $(\emptyset)$ también originados. La definición de sus valores se realiza mediante las siguientes fórmulas:

$$\sum_{i=1}^{n} R(i) = \sqrt{(J(i)_x - O_x)^2 + (J(i)_y - O_y)^2 + (J(i)_z - O_z)^2}$$

$$\sum_{i=1}^{n} \theta(i) = atan2\left(\sqrt{(J(i)_x - O_x)^2 + \left(J(i)_y - O_y\right)^2}, (J(i)_z - O_z)^2\right)$$

$$\sum_{i=1}^{n} \emptyset(i) = atan2\left((J(i)_x - O_x), \left(J(i)_y - O_y\right)\right)$$

La normalización llevada a cabo tiene la finalidad de menguar el impacto de la posición del usuario frente a la cámara, para luego almacenar cada joint mediante los valores correspondientes a su distancia al origen, y sus dos ángulos θ y $\emptyset$. De forma similar resulta necesario lidiar con el tamaño del usuario a través de otra normalización tomando en cuenta la distancia entre los joints de la cabeza J_{Head} y la mitad del cuerpo $J_{SpineMid}$, dichos valores varían proporcionalmente a la estatura. Esta normalización sólo se aplica sobre el valor correspondiente a las distancias radiales R de cada joint considerado antes de ser almacenados. La fórmula correspondiente a la normalización de la estatura es la siguiente:

$$\sum_{i=1}^{n} R_{norm}(i) = \frac{R(i)}{r_{HO}}$$

Culminados los pasos anteriores, los datos son almacenados en tres archivos diferentes, cada uno contiene información de un tipo de todos los frames correspondientes al gesto de los joints seleccionados.

En el Modo Prueba o Test, los autores plantean el empleo del algoritmo Nearest-Neighbor con DTW como propuesta de solución ante variaciones en el número de frames que pueden contener representaciones de un gesto, dicho algoritmo permite su clasificación a partir de comparaciones contra los existentes en la base de datos, realizando un alineamiento óptimo entre dos secuencias.

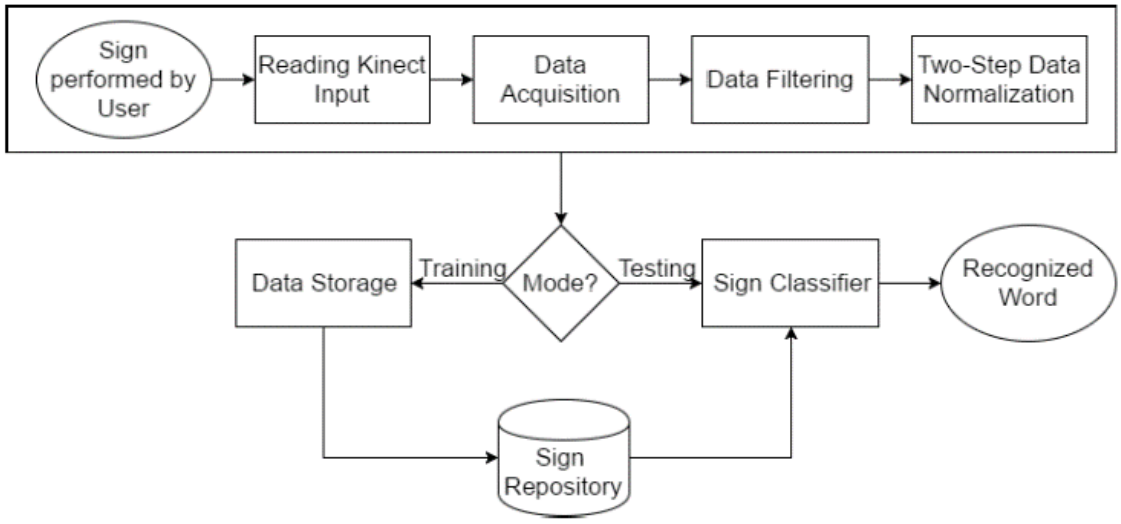


Figura 3.15 Descripción de la propuesta de Gkigkelos. Fuente: [51]

La propuesta fue desarrollada empleando Visual Studio 2015 Community Edition en .NET y la API correspondiente al Sensor Kinect provistos por Microsoft en su portal web.

Para la evaluación del sistema, emplearon 5 personas (2 hombres y 3 mujeres) con diferentes características físicas y sin conocimientos previos de la Lengua de Señas Griega. Cada uno de ellos realizó 15 signos cuatro veces (Las dos primeras en el centro de la escena, el tercero al lado derecho y la última en el lado posterior izquierdo) para entrenar al sistema. Los participantes fueron adiestrados mediante videos provistos por el Instituto de Política Educativa del país con los detalles suficientes para su correcto aprendizaje.

Los autores definieron 15 gestos, divididos en dos grupos, 6 conformados por aquellos que sólo necesitan la intervención de una mano y el resto el empleo de las dos. Como criterio de selección consideraron emplear signos que indican dirección por ser empleados en el día a día, y con las intenciones de poner a prueba el algoritmo de clasificación implementado, seleccionaron signos similares en su composición.

Los resultados de la evaluación son resumidos en la siguiente imagen donde se tienen diferentes configuraciones del aplicativo en búsqueda de la mejor.

Spherical Coordinates	Plain	Weighted	Median Filter	Reduced Joints	Reduced Samples	Center Only
Radial Distance (RD)	89.00%	86.00%	87.33%	89.00%	91.33%	94.67%
Zenith Angle (ZA)	81.67%	81.00%	80.00%	75.00%	81.33%	92.67%
Azimuth Angle (AA)	87.67%	87.67%	90.00%	88.33%	88.00%	98.00%
RD + ZA	89.00%	89.00%	89.00%	87.33%	88.33%	98.67%
RD + AA	**93.00%**	**92.67%**	**94.67%**	92.00%	**93.00%**	97.33%
ZA + AA	90.67%	89.67%	91.33%	89.00%	89.33%	98.00%
RD + ZA + AA	92.00%	92.33%	92.67%	**92.67%**	91.00%	**99.33%**

Figura 3.16 Evaluación de configuraciones para el sistema propuesto por Gkigkelos. Fuente: [51]

La primera columna revela diferentes consideraciones de los datos provenientes de las coordenadas esféricas. La primera fila presenta seis diferentes configuraciones para la clasificación:

1) Clasificación plana (Plain). Sólo emplea las coordenadas esféricas.
2) Clasificación ponderada (Weighted). Realiza la clasificación considerando las coordenadas esféricas con pesos. Los joints de las manos y codos tienen un peso del 90% y 10%, respectivamente
3) Clasificación de la mediana de filtrado (Median Filter). Emplean coordenadas esféricas con la aplicación de filtro de mediana.
4) Clasificación considerando menos joints (Reduced Joints). Sólo se usan seis joints (WL, HL, EL, WR, HR y ER) en vez de diez.
5) Clasificación considerando sólo dos individuos (Reduced Samples). Las pruebas se harán con una base de datos poblada con sólo datos obtenidos de dos individuos en la etapa de entrenamiento.
6) Clasificación considerando sólo la posición central (Center Only). La base de datos sólo contiene ejemplos de usuarios ubicados en la posición central de escena.

En la imagen anterior los mejores resultados son obtenidos al combinar la Distancia Radial (RD) y Ángulo Azimutal (AA) con el Filtro de Mediana (Median Filter) obteniendo un 94.67%, así también se logra una tasa de reconocimiento del 99.33% empleando las tres coordenadas esféricas (RD, ZA y AA) y considerando sólo el entrenamiento realizado desde la posición central de escena. Del cuadro además se puede

inferir que la clasificación plana ofrece mejores resultados que la segunda, indicando que los brazos juegan un rol importante como las manos debiendo tener la misma ponderación. Reduciendo la cantidad de joints a seis se logra menor resultado que el uso de diez como consecuencia de carecer del mismo nivel de detalle de la trayectoria de los gestos. La realización del Filtro de Mediana permite alisar los datos obtenidos del sensor, produciendo mejores resultados (94.67%). En la última columna se presentan mejores tasas de reconocimiento llegando al 99.33% en una configuración muy usada por otros autores, siendo el escenario más común en contextos de aplicación.

De los 15 gestos probados, los autores comentan haber logrado el 100% de tasa de reconocimiento para 10 de ellos, los otros cinco gestos presentaron inconvenientes durante las pruebas, cuyo detalle se puede observar en la siguiente imagen:

	book	key	white	paint	straight	middle	phone	hot	run
book	0,8					0,15			0,05
key		0,8	0,05				0,15		
white		0,05	0,9				0,05		
paint				0,8	0,05			0,15	
straight				0,05	0,9			0,05	

Figura 3.17 Probabilidades de errores en clasificación del sistema propuesto por Gkigkelos. Fuente: [51]

Como trabajo futuro plantean entrenar mayor número de signos por más individuos sin afectar la performance y tasa de reconocimiento. Así también, agregar el reconocimiento y seguimiento de expresiones faciales aprovechando las propiedades del sensor Kinect 2.

De Gkigkelos [51] se tomará en consideración el uso de herramientas software para la implementación de su propuesta, entre ellas se tiene el IDE Visual Studio 2015 Community Edition, bajo el lenguaje de programación .NET de forma conjunta con el SDK propio de la Kinect. Para la evaluación de su propuesta emplearon 5 personas (2 hombres y 3 mujeres) asegurándose de presentar diferentes características y sin conocimientos previos de la Lengua de Señas Griega. Cada persona realizó 15 signos cuatro veces en diferentes ubicaciones de la escena (2 en el centro, 1 en el lado derecho y el último en el izquierdo) para el entrenamiento del sistema. El medio de adiestramiento de las personas fue mediante videos provistos por el Instituto de Política Educativa de Grecia. Las evaluaciones de los autores bajo diferentes configuraciones arrojan que los mejores resultados (99.33%) de reconocimiento surge cuando se trabaja con la

combinación de tres coordenadas esféricas y entrenamiento realizado desde la posición central de escena.

3.4 Revisión de Técnicas para Tratamiento de Datos

El SDK desarrollado por Microsoft para el empleo del sensor Kinect carece de soporte para el reconocimiento de gestos, por lo tanto, el investigador deberá implementar el mecanismo que considere adecuado valiéndose de enfoques o definiciones de reglas sobre las posiciones de las partes del cuerpo. Sin embargo, debe superar limitantes como la variación en tamaño y ubicación de los usuarios, posturas, acercamiento, ángulo de visión, velocidad y destreza en el desarrollo del gesto, entre otros. [48] [38]

El enfoque más usado en el rastreo de manos es el uso de sensores de visión a partir de imágenes RGB o de profundidad, para posteriormente emplear las características extraídas en clasificadores estadísticos para tipificar y reconocer los gestos. En dicho punto, gran cantidad de métodos han sido propuestos e implementados: Scale-invariant Feature Transform (SIFT), Histogram of Oriented Gradients (HOG), Artificial Neural Networks (ANN), Support Vector Machines (SVM), Decision Trees (DT). [10]

Un inconveniente producto de la variedad de propuestas existentes es que la data de entrenamiento en la mayoría de casos no puede ser usado por otros como consecuencia de diferencias en calidad y representación, dificultando una comparación directa entre los resultados en performance y tasa de reconocimiento logrados. [52]

En el contexto, HMM y DTW han sido métodos de clasificación ampliamente implementados para el reconocimiento de voz y secuencias de movimiento, independientemente de la duración y variaciones de data.

HMM es catalogada como una técnica potencial en el proceso de reconocimiento de gestos, sin embargo, cuenta con limitaciones como: Necesidad de un gran set de entrenamiento y la incapacidad de ponderar características acordes a su importancia. Por lo anterior, diversos autores consideran necesario trabajar con otros métodos de forma conjunta, agregando técnicas de visión por computador para mejorar el rendimiento al extraer características geométricas de gestos y tomar ventaja de la ponderación. [41]

En la mayoría de las investigaciones de Reconocimiento de Lengua de Señas los métodos de clasificación requieren un entrenamiento explícito, resultando conveniente al causar bajo costo computacional durante el tiempo de ejecución de las pruebas. Sin embargo, el

agregar nuevos ejemplos incurre en complicadas fases de entrenamiento que deben iterarse repetidas veces. En contraparte, los Métodos Basados en Instancias[20] disipan la fase de entrenamiento costoso, pero pueden generar altos costes computacionales durante el proceso de clasificación, aletargando el reconocimiento en tiempo real de señales en grandes bases de datos de reconocimiento. [41]

A continuación, se muestran diferentes sistemas de reconocimiento de gestos, sus dataset, métodos de clasificación y tasas de reconocimiento.

Autor(es)	**Medio de Adquisición de Datos**	**Método de Clasificación**	**Dataset**	**Resultados**
Waldron y Kim	Guantes de datos y sensor Polhemus	BP Network, SON	14 ASL words	86%, 84%
Kadous	PoweGlove	IBL	95 Auslan word	80%
Vogler y Metaxas	Magnetic sensors and computer vision	HMM	53 ASL words	87.71% (CR)
Hernandez-Rebollar	DataGlove and accelerometers	A 3-level hierarchical classifier	26 ASL alphabet gestures	96.3%
Brashear	Accelerometers and hat-mounted camera	HMM	5 ASL gestures	90.48%
Holden	Colour-coded gloves	HMU	22 Auslan words	95%
Zhang	Multi-coloured gloves	TMDHMM	439 Chinese SL words	92.5%
Huang y Huang	Skin segmentation	HNN	15 Taiwanness SL words	91%

[20] Este tipo de entrenamiento almacena los ejemplos. La clasificación de un nuevo objeto requiere la extracción de los más parecidos para otorgarle la misma clasificación. También es conocido como Lazy Learning o Memory-Based Learning.

Holden	Skin segmentation with the snake algorithm	HMM	163 Auslan words	97% (SL)
Zieren y Krais	Multiple hypothesis	HMM	232 BSL words, 221 BSL words, 18 BSL words from 6 signers	99.3% (SD), 44.1% (SI), 87.8%
Starner y Pentland	Hat-mounted camera	HMM	40 ASL words	97% (CR)
Vogler y Metaxas	3 orthogonally placed cameras	HMM	53 ASL words	87.71% (CR)
Muñoz-Salinas	MLSHI	SVM	10 defined gestures	86.83%
Grzeszcuk	Stereo camera system	Statistical moments	6 defined gestures	96%
Lang, Block y Rojas	Kinect skeleton tracking	HMM	25 German SL words	97%
Ong	Kinect skeleton tracking	SP Trees	40 German SL words	55.4% (SI)
Kadir	Boosting 2 weak classifiers	A two-stage classifier	164 BSL words	92%
Wong y Cipolla	Motion gradient orientation images	Bayesin classifier	10 defined gestures	90%
Zahedi	Raw video	HMM	50 ASL words	82.8%
Cooper and Bowden	3 different classifiers	A 2-level classifier	164 BSL words	74.3%
Kelly	Hand postures from video	SVM	10 static gestures, 23 ISL letters	91.8%, 97.3%
Kim	Data glove	FMMNN	25 KSL words	25%
Lee	CyberGlove and Polhemus sensor	FMMNN	131 KSL words	80.1%
Yang	Motion and Skin Segmentation	TDNN	40 ASL words	93.42%

Kim	Fuzzy partitioning using speed	HMM	15 KSL sentences	94% (SL)
Vogler y Metaxas	MotionStar 3D tracking	PaHMM	99 sentences over 22 signs	84.9% (SL), 94.2% (WL)
Uebersax	Depth camera	ANMM	56 ASL words	97.8% (FS)
Feris	Colour camera wirh four flashes	Nearest-neighbour	ASL alphabet except 'J' and 'Z'	96% (FS)
Jerde	CyberGlove with sensors	Discriminant analysis	26 ASL letters	95% (FS)

Tabla 3.1 Porcentaje de reconocimiento de sistemas de terceros. Fuente: [41]

Según la tabla, el método de clasificación más recurrente es Hidden Markov Model (HMM) obteniendo unos porcentajes de reconocimiento interesantes. La selección de un algoritmo clasificador para el reconocimiento depende de las necesidades de la aplicación, por ejemplo, FSM (Máquina de Estados Finito) es simple y fácil de implementar, pero de lento procesamiento, mientras DTW es un método ventajoso en aplicaciones de reconocimiento continuo. [53]

En Carmona [54] realizan comparaciones desde diferentes criterios (Porcentajes de reconocimiento, sensibilidad a la cantidad de muestras de entrenamiento, parámetros óptimos y tiempos de procesamiento) a los métodos Hidden Markov Model (HMM) y Dynamic Time Warping (DTW) para determinar cuál presenta mayor performance y tasa de reconocimiento, empleando un conjunto de gestos conformados por data obtenida desde el sensor Kinect con OpenNI mediante Skeleton Tracking, entrenando 2500 ejemplos por cada método. Usaron 5-state HMMs y 3-Nearest Neighbour Classifier como métodos de clasificación de gestos con la distancia obtenida por DTW. A continuación, se muestra la tabla con los resultados, donde se tienen el mínimo y máximo promedio de reconocimiento:

	Mínimo	**Máximo**	**Promedio**
HMM	92.8%	99.2%	96.46%
DTW	97.2%	100%	98.84%

Tabla 3.2 Promedios de reconocimiento. Fuente: [54]

Otro aspecto consistió en evaluar los tiempos de procesamiento requeridos para el entrenamiento y reconocimiento. El tiempo necesario para entrenar HMM con 50 ejemplos de cada clase de gesto es 17.3 microsegundos (ms), mientras que DTW no requiere explícitamente una etapa de entrenamiento. HMM tarda 3.6 ms en clasificar un gesto a partir de secuencias de entrenamiento o vectores de características, y DTW emplea en promedio 6 microsegundos. Sólo se necesitan 3 muestras o ejemplos de DTW para lograr una tasa de error similar a la obtenida con HMM con 50 muestras.

Las conclusiones mostradas en Carmona [54] demuestran que DTW logra mejor perfomance en comparación con HMM en el reconocimiento de gestos, este último requiere mayor entrenamiento para obtener una performance similar al primero. Sin embargo, el tiempo del proceso reconocimiento DTW depende directamente a la cantidad de muestras almacenadas en la base de datos.

En Ibañez [48] presentan una herramienta software la cual facilita el desarrollo de aplicaciones que interactúan por medio de gestos basados en Kinect. GRTool faculta al desarrollador construir un conjunto de gestos de entrenamiento y provee soporte para construir clasificadores que identifiquen gestos similares a los entrenados. GRTool permite el entrenamiento de diferentes técnicas de Machine Learning para el reconocimiento de gestos: Dynamic Time Warping, Hidden Markov Models y Procrustes Analysis[21]. El objetivo de Ibañez [48] consistió en evaluar dos aspectos: La precisión de las técnicas de reconocimiento y los esfuerzos que requiere desarrollar una aplicación con y sin GRTool, obteniendo como resultados 97% de precisión de los gestos evaluados y una reducción del 67% del esfuerzo de desarrollo en términos de líneas de código. En esta investigación recomiendan usar HMM si lo deseado es alta exactitud de reconocimiento en términos de calidad de gesto imitado, mientras que DTW o Procrustes deben ser usados si se desea obtener más información sobre cómo el gesto está siendo reconocido. Para obtener los resultados empleando la técnica HMM requieren definir previamente la

[21] Esta técnica busca la alineación óptima entre dos trayectorias de partes del cuerpo, aplicando una serie de transformaciones matemáticas. Las transformaciones son Centrar, Normalizar y Rotar. Las primeras buscan reducir el impacto de las variaciones físicas de los usuarios, mientras el último consiste en rotar las trayectorias hasta obtener una mínima distancia entre ellas. Para hallar dicho valor, usa como criterio la diferencia de cuadrados entre los puntos de las trayectorias. Como resultado se obtiene la distancia o diferencia de las dos trayectorias, pudiéndose emplear de la misma manera que DTW.

cantidad de clústeres a utilizar, trabajando con 5, 10, 15 y 20 a fin de determinar el impacto de la cantidad de clústeres en la tasa de reconocimiento.

Para cada conjunto de gestos entrenados se obtuvieron los siguientes resultados:

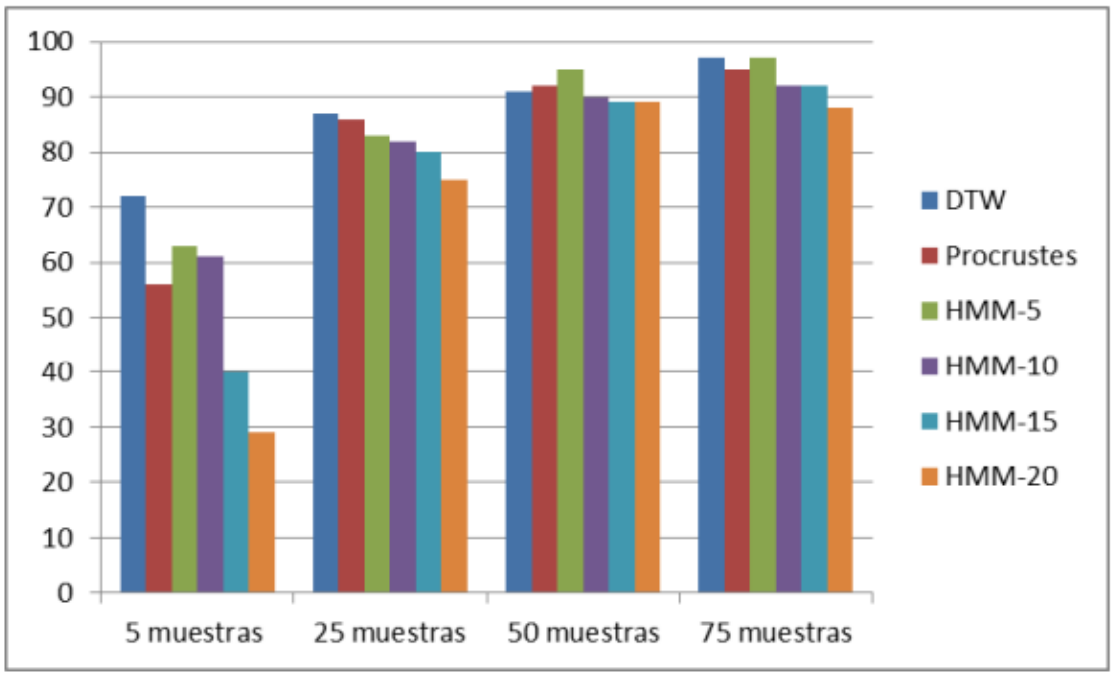


Figura 3.18 Gráfico comparativo de la cantidad de gestos reconocidos por cada técnica variando la cantidad de muestras utilizadas para el entrenamiento realizado por Ibañez. Fuente: [48]

Según la imagen, aumentar la cantidad de ejemplares utilizados en el entrenamiento influye de forma positiva en la precisión del reconocimiento de todas las técnicas. Así también, la cantidad de clústeres HMM degrada la tasa de reconocimiento a partir de 50 muestras como consecuencia de incrementar o exigir la fidelidad del gesto a reconocer, por lo cual la imitación y reconocimiento de un gesto se torna más difícil de alcanzar al tener que pasar éste por cada uno de los clústeres definidos, en este punto es necesario establecer un balance entre la exactitud de la técnica y la usabilidad del sistema.

En Xu [55] realizan una evaluación entre DTW y HMM mediante un proyecto open-source nombrado Speech Recognition System Based on Matlab, para demostrar cuál es más óptimo en el reconocimiento de gestos. La evaluación se realizó sobre el reconocimiento de diálogos y no de gestos, no obstante, el autor generalizó los resultados obtenidos como válidos. La imagen que resume las conclusiones obtenidas se muestra a continuación.

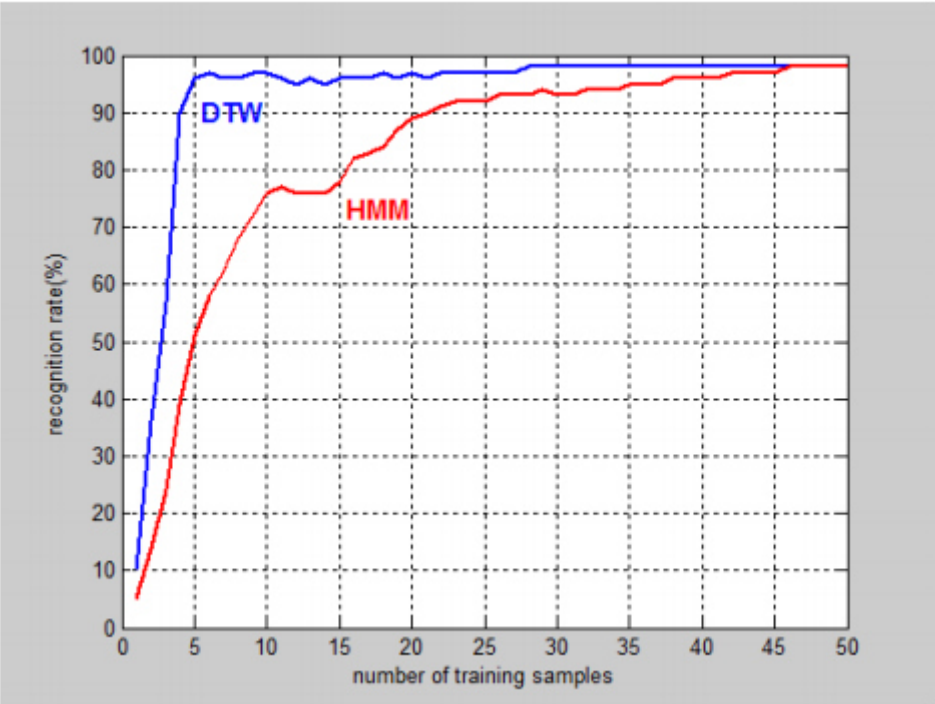


Figura 3.19 Comparación de performance entre DTW y HMM con diferente número de ejemplos de entrenamiento. Fuente: [55]

De la figura anterior se deduce que DTW provee aproximadamente un 90% de reconocimiento después de sólo cuatro ejemplos de entrenamiento, y logra un 98% después de 27, mientras que HMM requiere unos 45 ejemplos de entrenamiento para asemejar su tasa de reconocimiento.

DTW otorga una adecuada e interesante tasa de reconocimiento a una mínima cantidad de muestras de entrenamiento. Tomando en consideración las evaluaciones y resultados mostrados en las diversas partes de este capítulo, se opta por elegir DTW como método de clasificación para el reconocimiento de gestos en este trabajo.

CAPÍTULO IV: DESARROLLO DE LA SOLUCIÓN PROPUESTA

El presente capítulo describe el desarrollo del sistema propuesto. El objetivo de la solución es reconocer en tiempo real los gestos realizados por una persona frente al sensor Kinect considerando el desplazamiento de brazos y manos. Dicho sensor será el encargado de captar los datos de profundidad para su posterior procesamiento.

4.1 Esquema General de Desarrollo

Un sistema se define como un conjunto de entidades que se encuentran relacionadas entre sí con la finalidad de contribuir al logro de un objetivo, estando compuesto de entradas, procesos y salidas.

La metodología[22] para el desarrollo de la tesis estará basada en Lopez-Ludeña [56] siendo una adaptación de Diseño Participativo[23], resultando ser uno de los enfoques más usados del Diseño Centrado en el Usuario que sigue el estándar de Diseño Centrado en el Usuario para Sistemas Interactivos (ISO 9241-210, 2010). Se caracteriza por involucrar a los stakeholders[24] desde el proceso de diseño buscando como producto final un software usable y capaz de satisfacer las necesidades por las cuales fue desarrollado. Prioriza las necesidades y juicios de los involucrados, define aspectos básicos para la etapa de pruebas (Cantidad de personal y tiempo necesario), invita al análisis de tasas y porcentajes obtenidos, así como conocer la satisfacción de los usuarios involucrados. Por último, en Lopez-Ludeña [56] se dan dos escenarios de aplicación, uno de ellos se similar a lo que se planea lograr con los alcances definidos en esta tesis, y el otro con los deseados como trabajo futuro, adecuándose fácilmente al presente trabajo y acaparando aspectos

[22] Busca asegurar que el software alcance los requerimientos de calidad (Performance, tiempo y costo), y usabilidad, lo cual conllevaría a una mejor aceptación del público al cual está dirigido, en este caso quienes empleen Lengua de Señas para comunicarse.

[23] También llamado Diseño Cooperativo.

[24] Personas interesadas.

considerados necesarios para el correcto desarrollo de la propuesta. Es gracias a los puntos mencionados que se decide acoger dicha metodología como guía.

La metodología se compone de cuatro fases: Análisis de Requerimientos, Recopilación de Datos, Adaptación de Tecnología, y Evaluación del Sistema. A continuación, se detalla cada una de ellas:

1. Análisis de Requerimientos: En esta etapa se analizan y recogen los requerimientos del público objetivo (Personas que emplean Lengua de Señas para comunicarse) y características del contexto (Área de atención al público) en relación al sistema propuesto cuya finalidad es facilitar la comunicación e interacción entre oyentes y no oyentes.
    1.1. Requerimientos de Usuario: Comprende entrevistas con los usuarios finales, para conocer los requerimientos y definir el alcance.
    1.2. Requerimientos Técnicos: Tiene como objetivo definir la calidad deseada del producto (Porcentaje de reconocimiento esperado y velocidad de procesamiento). Para lograr satisfacerla es necesaria la revisión de diferentes técnicas, analizarlas, compararlas y/o combinarlas para definir la más adecuada.
2. Recopilación de Datos:
    2.1. Selección de Escenarios y Área de Aplicación: Mediante entrevistas con los usuarios finales se puede determinar el área donde el software será aplicado.
    2.2. Definición de Términos a Traducirse: Comprende el establecimiento de expresiones, palabras, frases y oraciones que el software deberá traducir según el alcance. Idealmente esto se obtendría del área de aplicación, donde los usuarios oyentes y no oyentes pueden facilitar las expresiones comunes.
    2.3. Traducción de las Oraciones Definidas: Todos los términos definidos en el paso anterior deberán ser traducidos a la Lengua de Señas objetivo por personas conocedoras del tema.
    2.4. Grabación de Video: Las oraciones o expresiones definidas deberán ser representadas por expertos en Lengua de Señas para ser almacenadas.
3. Adaptación de Tecnología: En este punto se describen los componentes software y aspectos funcionales de la propuesta a desarrollar. Por ejemplo, en [56] se describen dos sistemas, uno de ellos permite al usuario verificar los gestos con los

que cuenta, seleccionándolos y viendo su representación mediante un avatar[25], además permite agregar nuevos gestos que se almacenarán para su posterior empleo.

4. Evaluación del Sistema:
   4.1. Diseño del Campo de Evaluación: Define aspectos necesarios como el número de días a realizar la evaluación y la cantidad de personas a involucrarse (Éstos pueden ser usuarios que empleen o conozcan Lengua de Señas, o sólo oyentes), además realizar evaluaciones estableciendo diferentes escenarios.
   4.2. Mediciones Objetivas: Las mediciones deben contener información objetiva obtenida por el sistema, pueden incluir los siguientes aspectos: Porcentaje o tasa de reconocimiento de gestos, velocidad de traducción, tiempo promedio de traducción, tiempo promedio de conversión de texto a voz, tasas de reconocimiento al emplear diferentes técnicas, número de expresiones con las que se trabaja, entre otros.
   4.3. Mediciones Subjetivas: Estas mediciones pueden ser obtenidas a partir de cuestionarios resueltos por los usuarios finales. Permiten evaluar diferentes figuras del sistema dando alguna puntuación. En el caso de los formularios destinados a los usuarios que empleen lengua de señas para comunicarse se debe tener en consideración aspectos como contar con intérpretes que puedan solventar dudas y atender incidentes en todo momento.

El esquema general de la propuesta de Lopez-Ludeña [56] fue aplicado al servicio de atención de personal que emplea Lengua de Señas Española (LSE) en dos escenarios que se denotan en la fase Adaptación de la Tecnología. El primero, Generador de Voz desde Lengua de Señas Española (Speech Generator from Spanish Sign Language) siendo similar al objetivo propuesto en esta tesis: A partir de un gesto realizado por el usuario el sistema procesa y obtiene como resultado la traducción correspondiente, permitiendo además agregar nuevos gestos a la base de datos y representarlos mediante un avatar para su validación visual por el encargado del entrenamiento. Y la segunda, Voz a Lengua de

[25] Entiéndase como la representación de una persona mediante animaciones 2D o 3D, en este caso empleados con la finalidad de simbolizar gestos según los parámetros que recibe.

Señas Española (Speech into Spanish Sign Language) consiste en traducir el diálogo verbal en palabras, para posteriormente convertirlas en secuencias de signos que serán representadas por un avatar.

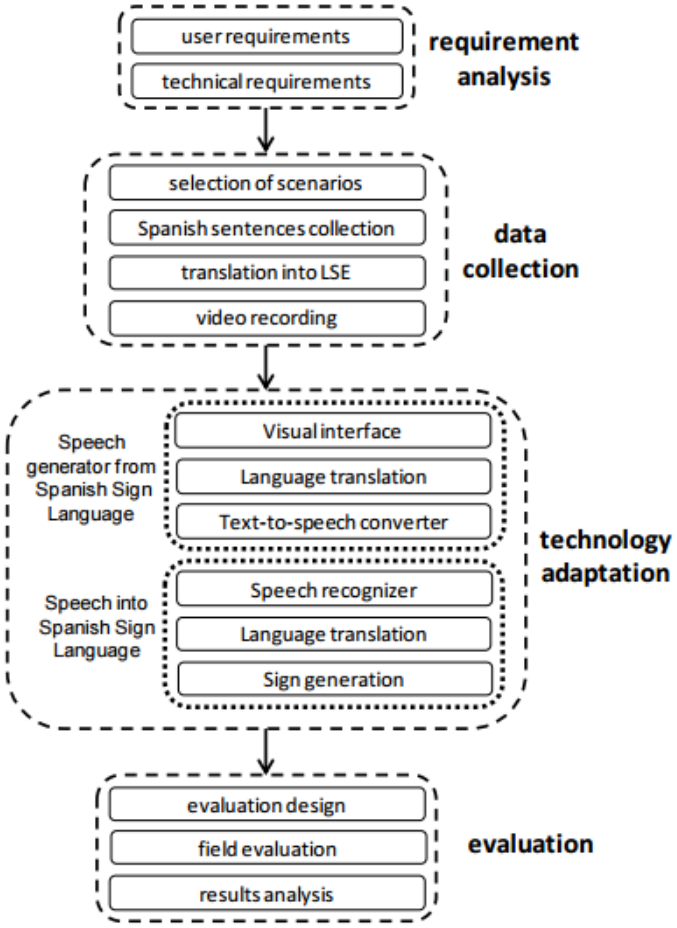


Figura 4.1 Fases de la metodología propuesta por Lopez-Ludeña. Fuente: [56]

4.2 Análisis de Requerimientos

Como se mencionó en el punto 1.1 en nuestro país en el año 2013 se contaron 532,209 personas que presentaban dificultades para oír, de ellos un 55.7 por ciento debía recurrir a alguna forma especial de comunicación (Gestos, lectura de labios, lenguaje de señas, lápiz y papel, entre otros), convirtiéndose en una limitante en la educación, oportunidades laborales, aceptación familiar, entre otros. [Anexo 02] [Anexo 04] [Anexo 06]

Como consecuencia resulta conveniente y necesaria la aplicación de tecnologías capaces de facilitar la comunicación, permitiendo la traducción automática de los gestos realizados a palabras, o viceversa[26].

Por tal motivo se procede a definir el contexto de aplicación al cual se orientará el desarrollo y entrenamiento del software. La aplicación tiene por objetivo servir de ayuda en los centros de atención presencial del público. En dicho contexto se identifican a dos tipos de usuarios recepcionistas: Las personas no oyentes quienes emplean lengua de señas para comunicarse (Pudiendo ser de cualquier edad y teniendo que contar con un adecuado conocimiento de los gestos) y además los oyentes quienes se valdrán de los resultados procesados por el software para entender y comunicarse con la contraparte.

4.2.1 Descripción General del Servicio otorgado en el Contexto de Aplicación

El contexto sobre el cual se orientará el desarrollo y entrenamiento del software está enfocado al área de atención al público en general. Sin embargo, para la realización de pruebas con respecto a temas técnicos como la tasa de reconocimiento del software, se vio adecuado tomar como referencia el área de recepción de los planteles de educación básica regular. Se observan dos escenarios: Una persona no oyente quien con el apoyo del software labora como recepcionista en la entidad, de tal forma puede interactuar con el personal que arriba a la institución para solventar sus dudas o realizar algún trámite. Mientras el segundo escenario consta de un usuario oyente recepcionista quien mediante el software puede entender los gestos realizados por algún apoderado o padre de familia que emplea lengua de señas para comunicarse.

4.2.2 Requerimientos de Usuario

Para conocer los requerimientos de usuarios se realizaron pequeñas entrevistas a colaboradores involucrados en contextos relacionados al definido en el punto anterior.

Se intervino a dos trabajadores de la Institución Educativa Nº 20915 – Pucará (Ubicado en Cucuya, distrito de Santo Domingo de los Olleros, provincia de Huarochirí) bajo el

[26] La traducción de palabras a gestos no está contemplada en el alcance de esta tesis.

cargo de Secretaria y Director, quienes se hacen cargo de la atención de los padres de familia y público en general, así como de llevar adelante los trámites que se soliciten.

- Proceso

  Cuando llega un padre de familia o tutor a la institución, la persona encargada lo recibe, consulta sus datos personales y el motivo de su visita. En caso el tutor es no oyente y el recepcionista desconozca el significado de los gestos entonces la comunicación entre ambas partes usualmente se realiza mediante papel o pizarra en lenguaje escrito.

  Una vez definido el trámite a realizar, por ejemplo, si es matrícula se procede a explicar al tutor la documentación, requisitos y los formularios a completar. Si el motivo de reunión es por indisciplina del alumno, se dan al tutor las recomendaciones para lograr un buen comportamiento.

  Si el proceso ha culminado con éxito, se consulta por si el tutor desea realizar o conocer algo más, caso contrario el recepcionista invita al paciente a retirarse.

- Definición del tipo de usuario

  Se identifican los siguientes tipos de usuario:

  a) Tutor: Generalmente son los Padres de Familia, quienes acuden al plantel educativo para ser atendidos y realizar algún trámite o proceso. El género y edad de éste es indistinto, pero debe tener relación con algún potencial educando o estudiante ya registrado en la institución.
  b) Recepcionista: Es el personal quien labora en la institución educativa y además es encargado de la recepción de las personas que llegan para realizar algún trámite o informarse.

  Estos actores pueden ser oyentes o no oyentes, si alguno de ellos es no oyente, entonces el software actuará como medio de apoyo para facilitar la traducción de los gestos correspondientes a la lengua de señas.

- Información intercambiada con frecuencia entre los usuarios definidos

  Entre estos ítems se tienen los siguientes:

  a) Solicitud de información acerca de: Proceso de matrícula y disciplina del menor en el plantel.
  b) Datos del Tutor: Nombres, número y nombre de lo(s) hijo(s).

c) Recomendaciones y requisitos brindados al Tutor para completar la documentación para llevar a cabo algún proceso y sugerencias para una adecuada corrección del mal comportamiento de su menor hijo.

4.2.3 Requerimientos Técnicos

En este punto se especifican aspectos técnicos que deberán ser satisfechos por el producto software. Entre estos se han establecido los siguientes:

- Considerando el Estado del Arte revisado, el sistema deberá lograr una tasa de reconocimiento superior al 95% para el conjunto de gestos dentro del alcance definido.
- Los gestos ejecutados deberán ser reconocidos en tiempo real procurando mantener una comunicación fluida y constante entre las partes.
- El sistema deberá trabajar sin mayores inconvenientes en los escenarios definidos.
- El usuario podrá agregar nuevos gestos al sistema para su posterior reconocimiento, sin degradar la performance y velocidad de procesamiento.

4.3 Recopilación de Datos

Esta fase se compone de los siguientes pasos:

4.3.1 Selección de Escenarios y Área de Aplicación

Como se mencionó anteriormente, el software estará orientado a servir de intermediario en áreas de atención al público de forma genérica, permitiendo facilitar la comunicación entre las personas oyentes y no oyentes.

4.3.2 Definición de Términos a Traducirse

Acorde a la información obtenida y mostrada en el punto 4.2.2, se identificaron un conjunto de términos comúnmente empleados en la interacción entre el recepcionista y tutor, realizándose además su clasificación y ordenamiento por fases según ocurre un proceso normal de atención al público.

Cabe resaltar que no todos los términos existentes en nuestro lenguaje tienen traducción a la Lengua de Señas Peruana, para conocer dicha realidad y definir el listado final de

expresiones se contó con el invaluable apoyo de Otto Ángel quien es una persona no oyente, pero además conoce y dicta talleres de Lengua de Señas Peruana en diferentes instituciones.

FASE	TÉRMINO	FRASE
Saludo	Hola	
	Bienvenido	
	Director	
	Curso	
	Docente	
	Siéntese	
	Espere	
Definición del Trámite	Indisciplina	
Procesamiento de Solicitud		¿Qué grado?
Entrega de la Solicitud	Terminado	
Despedida		No se encuentra

Tabla 4.1 Términos a traducirse. Fuente: Elaboración propia.

Teniendo definido el listado de expresiones que el software deberá traducir, se procederá a ilustrar mediante fotos la realización de cada una de ellas:

1) Hola

Figura 4.2 Representación de la Seña: Hola. Fuente: Elaboración propia

2) Bienvenido

Figura 4.3 Representación de la Seña: Bienvenido. Fuente: Elaboración propia

3) Director

Figura 4.4 Representación de la Seña: Director. Fuente: Elaboración propia

4) Curso

Figura 4.5 Representación de la Seña: Curso. Fuente: Elaboración propia

5) Docente

Figura 4.6 Representación de la Seña: Docente. Fuente: Elaboración propia

6) Siéntese

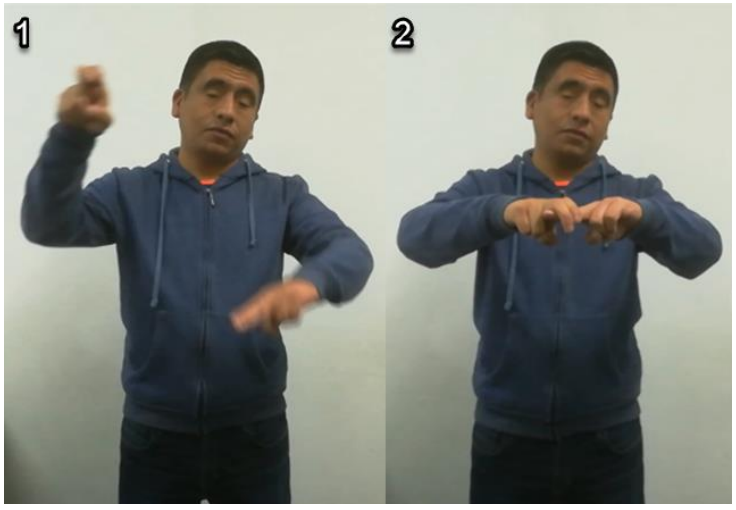


Figura 4.7 Representación de la Seña: Siéntese. Fuente: Elaboración propia

7) Espere

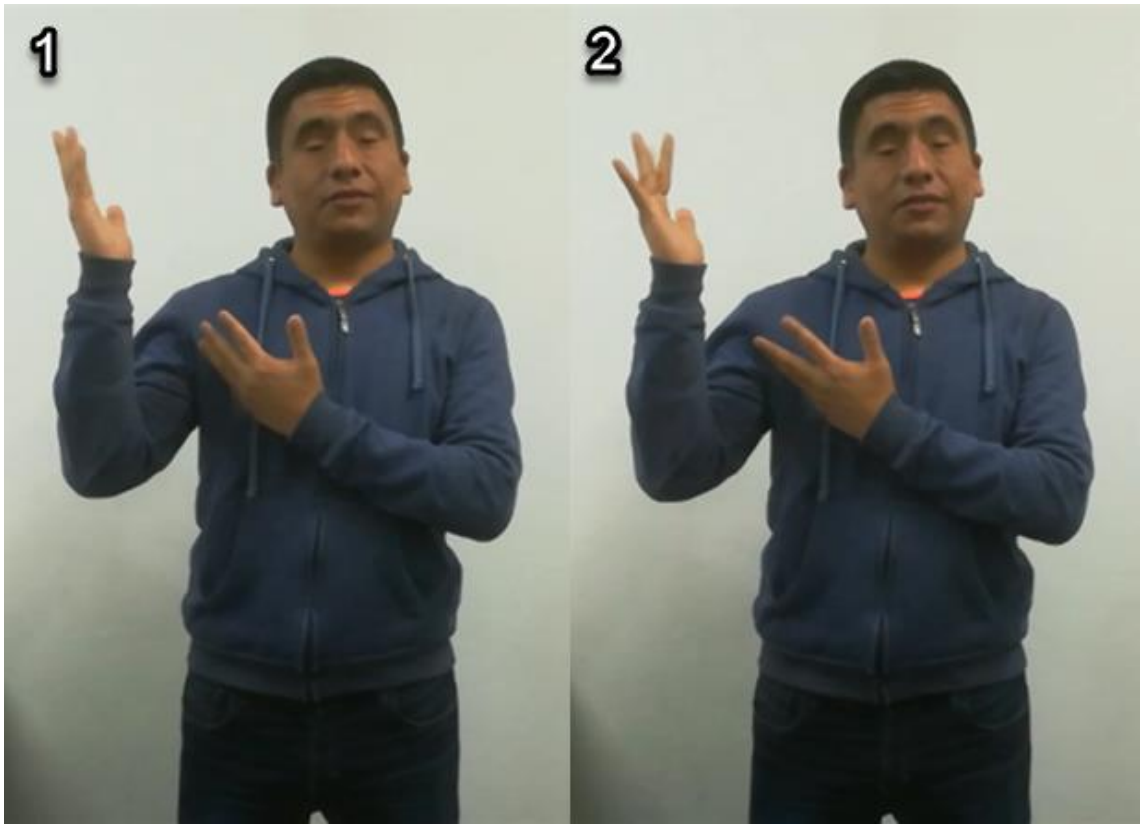


Figura 4.8 Representación de la Seña: Espere. Fuente: Elaboración propia

8) Indisciplina

Figura 4.9 Representación de la Seña: Indisciplina. Fuente: Elaboración propia

9) ¿Qué grado?

Figura 4.10 Representación de la Seña: ¿Qué grado? Fuente: Elaboración propia

10) Terminado

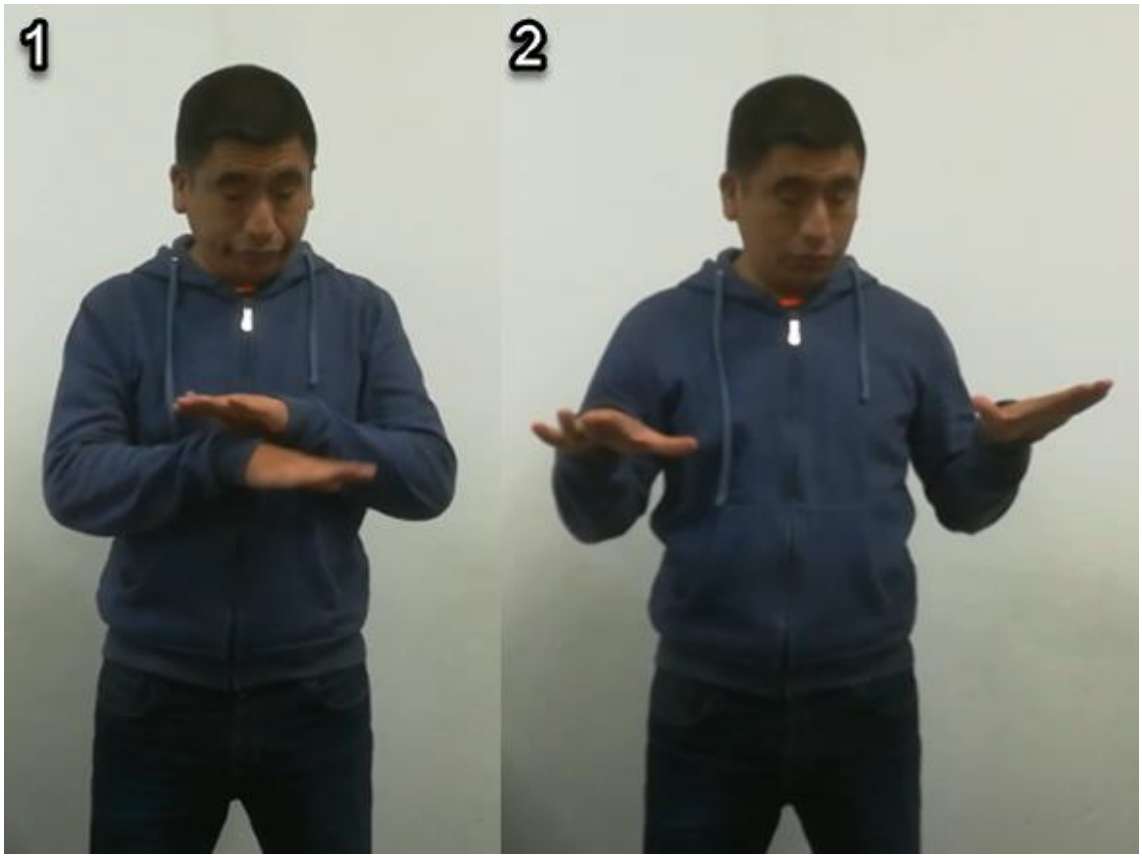


Figura 4.11 Representación de la Seña: Terminado. Fuente: Elaboración propia

11) No se encuentra

Figura 4.12 Representación de la Seña: No se encuentra Fuente: Elaboración propia

4.3.3 Entrenamiento del Sistema

El personal encargado de entrenar el aplicativo estará compuesto por personas diestras en el empleo de Lengua de Señas Peruana, preferentemente quienes la usen como medio de comunicación o se desempeñen como intérpretes.

Teniendo en consideración el Estado del Arte revisado y con la finalidad de evaluar la tasa de reconocimiento del aplicativo dado un mínimo entrenamiento por cada gesto, se ha decidido realizar esta etapa con la ayuda de sólo una persona no oyente quien emplea Lengua de Señas Peruana como medio habitual de comunicación. Dicha persona estará ubicada en la parte central del área de grabación y realizará cinco veces cada gesto definido.

4.4 Diseño del Sistema Propuesto

Este punto describe los módulos o componentes software relacionados entre sí con el fin de cumplir el objetivo planteado.

Este punto se divide en 5 partes y para un mejor entendimiento se describe en la siguiente imagen el flujo general de los componentes a implementarse bajo el alcance especificado.

4.4.1 Adquisición de Datos

La obtención de datos y su posterior procesamiento requiere del empleo de un adecuado software y hardware, los cuales dependen de una correcta captura de datos, por tal motivo las características técnicas y pasos necesarios para la instalación del sensor Kinect se detallan en [Anexo 07].

Como input se tendrán los datos de flujos de videos (RGB y profundidad) capturados por el sensor, estos son capaces de entregar hasta 30 frames por segundo. El aplicativo mostrará al usuario posicionado frente a la cámara su Skeleton Tracking[27].

El software estará habilitado para el seguimiento y reconocimiento de los movimientos de una persona a la vez.

[27] A partir de ahora se usarán los términos Rastreo del Cuerpo para referirnos a ello.

El usuario tendrá la posibilidad de activar la opción Grabar[28], caso contrario el sistema solicitará el siguiente frame y refrescará el video mostrado en pantalla perdiendo la información del frame anterior.

[28] Permite indicar al sistema que debe capturar y almacenar los frames que sean generados por el usuario posicionado frente a la cámara, éste puede estar realizando algún gesto.

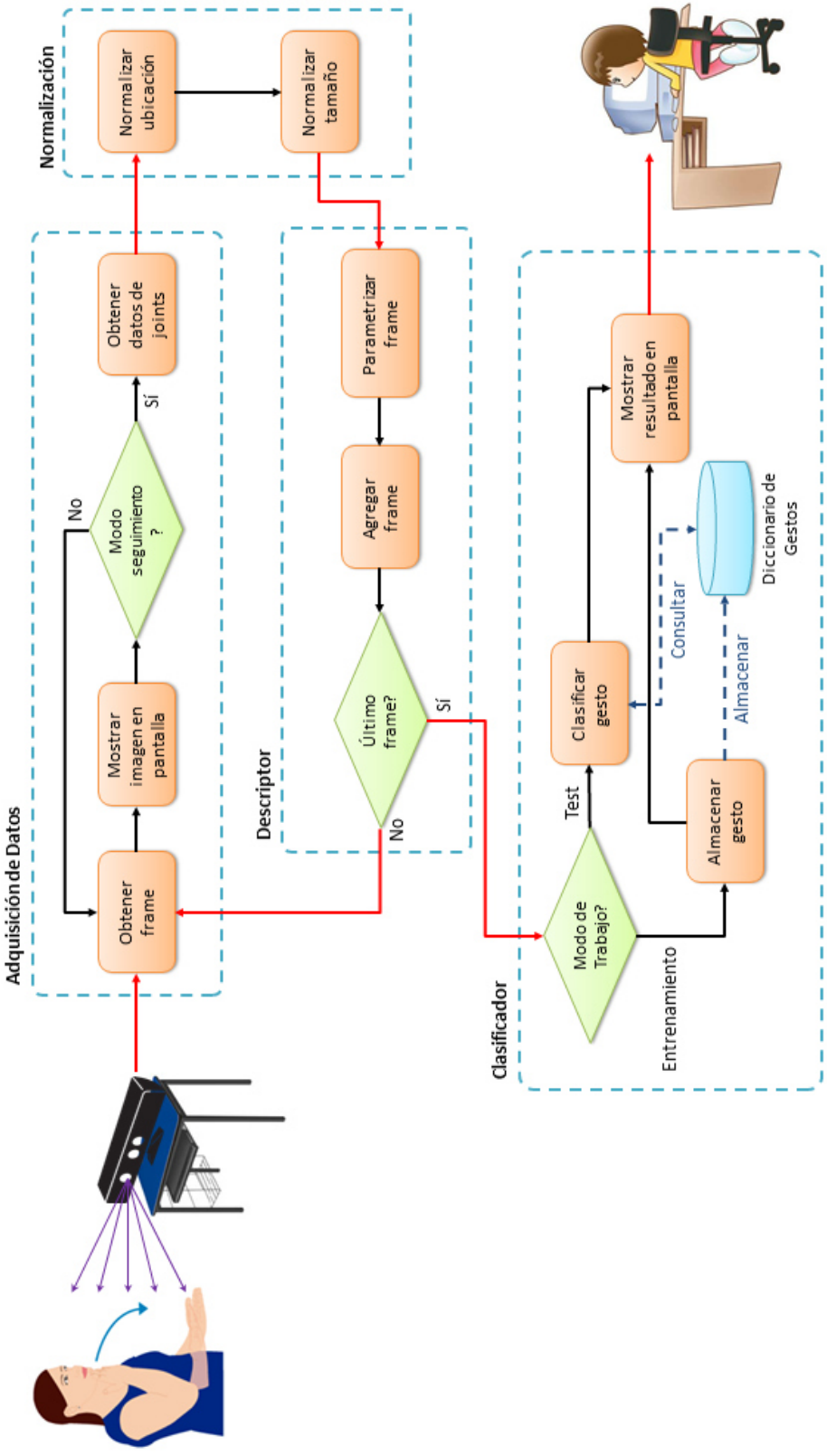


Figura 4.13 Diagrama de flujo del sistema propuesto. Fuente: Elaboración propia.

Figura 4.14 Captura de imagen RGB y seguimiento del cuerpo humano de forma simultánea. Fuente: Elaboración propia.

Si la opción Grabar se encuentra activada, el sistema mostrará una ventana solicitando la acción a realizar (Reconocer Seña o Agregar Seña) y la cantidad de segundos antes de empezar la grabación. Si el usuario indica Agregar Seña, el sistema solicitará además el Nombre de la Seña que va a realizar.

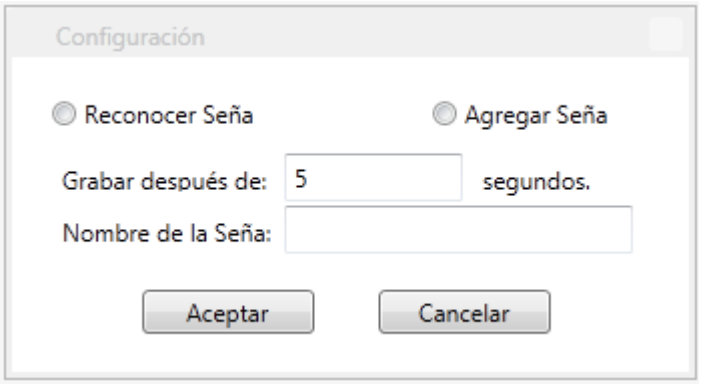


Figura 4.15 Ventana de configuración del sistema propuesto. Fuente: Elaboración propia.

Una vez confirmados los valores ingresados, el sistema empezará el conteo regresivo para obtener los datos de las posiciones X, Y y Z de los joints de interés en cada frame.

Figura 4.16 Conteo regresivo. Fuente: Elaboración propia.

4.4.2 Definición de Joints de Interés

El SDK definido permite detectar y realizar el seguimiento de 20 joints (Articulaciones) del cuerpo humano (Figura 2.4), algunas resultan irrelevantes para el alcance definido por lo cual en este trabajo se emplearán sólo aquellos considerados importantes o necesarios.

Tal como se concluyó a partir del Estado del Arte, el aplicativo rastreará los datos provenientes principalmente del desplazamiento de 4 joints: Ambas manos y codos; gracias a que los joints restantes, en la performance de todos los gestos, pierden significancia al mantenerse constantes conllevando a redundancia y costos computacionales innecesarios.

Sin embargo, tal como se mencionó anteriormente, en las etapas de procesamiento siguientes resulta necesario contar con data de joints adicionales para hacer frente a problemas causados por variaciones en las características físicas de usuarios y sus ubicaciones frente al sensor. Por tal motivo, también se considerarán los joints correspondientes a la Cabeza y Espina[29] (Ver figura anterior), estos dependen directamente de las características físicas y posición del usuario, generalmente se mantienen visibles al sensor durante la realización de gestos lo cual facilita su obtención y trabajar con ellos sin mayor inconveniente.

[29] También llamado Torso.

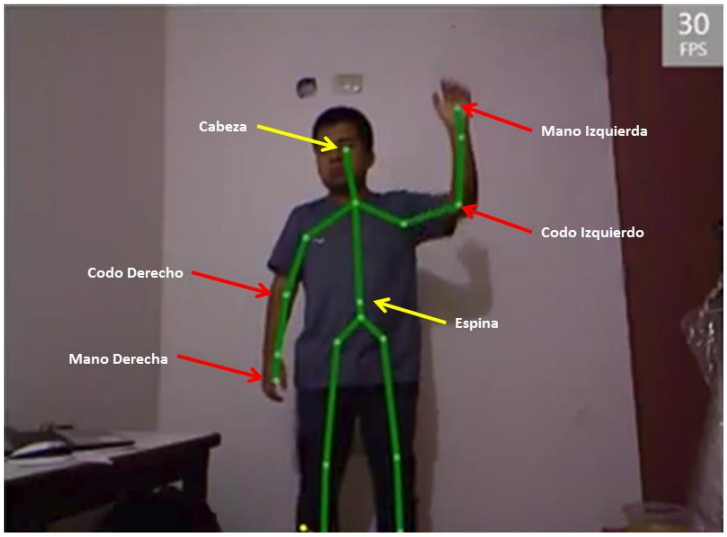


Figura 4.17 Ubicación de joints de interés en el rastreo del cuerpo. Fuente: Elaboración propia.

4.4.3 Normalización

El Estado del Arte evaluado en Ibañez [48] presenta limitaciones al hacer frente a las variaciones de ubicación y características físicas del usuario en el área de detección, por lo tanto, un sistema robusto debe ser capaz de normalizar la data obtenida para reducir o evitar errores de reconocimiento.

Esta normalización está compuesta por dos fases:

4.4.3.1 Normalización de la Ubicación del Usuario

El usuario puede ubicarse en diferentes lugares del Área de Detección[30], por lo que la data variará acorde a ello. Es decir, una pequeña alteración causará diferentes valores X, Y y Z como se observa a continuación.

[30] Se denomina así al campo físico en el cual el sensor es capaz de realizar detecciones.

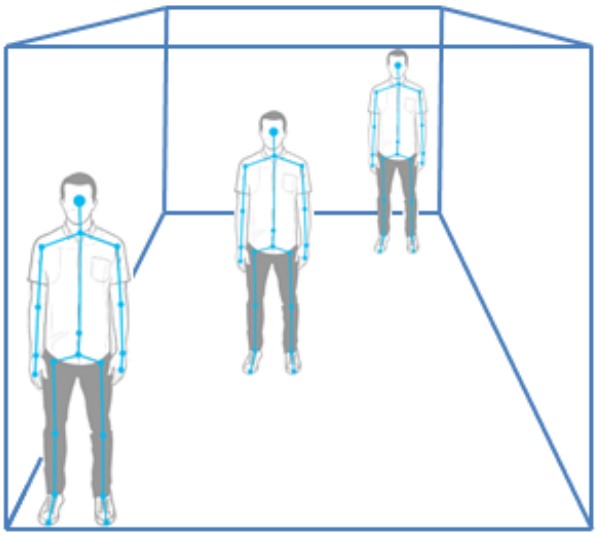


Figura 4.18 Variaciones en la ubicación del usuario. Fuente: Elaboración propia.

Para abordar ello se realizará la normalización a partir de la posición del joint del Torso (Espina), este se mantiene constante durante el tiempo de captura de frames dependiendo directamente de la ubicación y características físicas del usuario.

Considerando lo anterior, se define al Torso como el origen para trabajar convenientemente con coordenadas esféricas[31] tal como se muestra en la siguiente figura.

Se define la distancia ***r*** representada por ***d*** y el vector entre el Torso y el correspondiente joint de interés (Codo o Mano), así también $\boldsymbol{\theta}$ y $\boldsymbol{\varphi}$ son los ángulos que describen la dirección 3D del vector representado en la siguiente imagen.

[31] En matemáticas, constituye otra generalización de las coordenadas polares del plano, a base de girarlas alrededor de un eje. Se define de la siguiente forma: La coordenada *radial r* es la distancia al origen. La coordenada *polar* θ es el ángulo que el vector de posición forma con el eje Z. Y la coordenada *acimutal* φ formado por el ángulo que la proyección sobre el plano XY forma con el eje X.

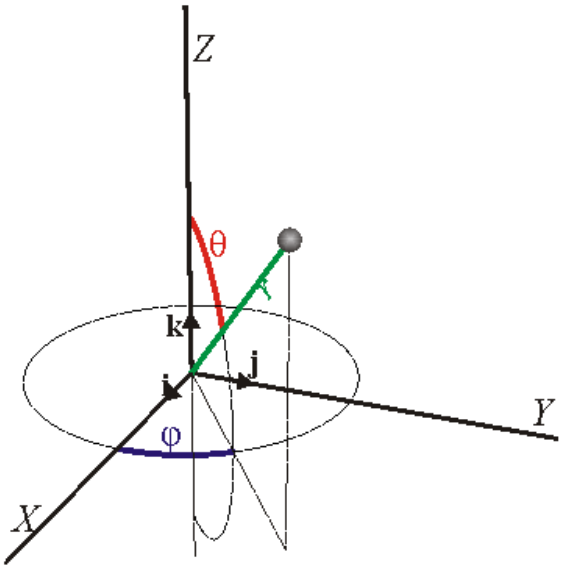
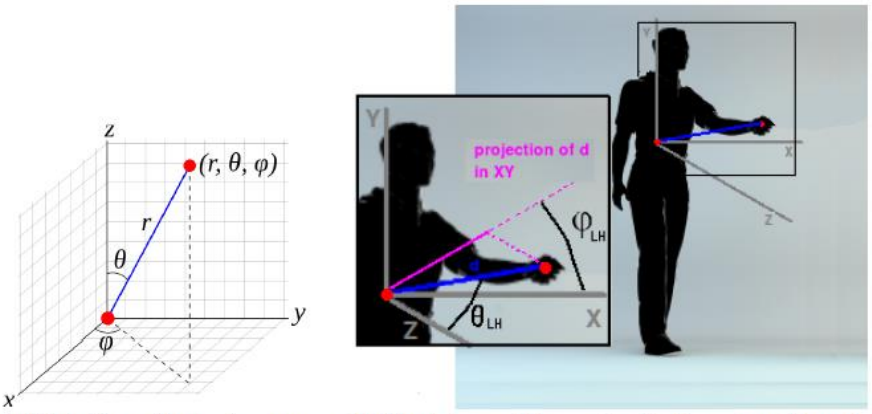


Figura 4.19 Coordenadas esféricas en la propuesta de Capilla. Fuente: [31]



(a) Definition of (r, θ, φ) as commonly used in physics: radial distance r, polar angle θ, and azimuthal angle φ. [26]

(b) Equivalence of these values in the system. Example case of the LH joint

Figura 4.20 Empleo de las coordenadas esféricas en el sistema propuesto por Capilla. Fuente: [31]

Considerando el conjunto de joints *J* definidos para el desarrollo, se tiene:

CD = Codo Derecho, CI = Codo Izquierdo, MD = Mano Derecha, MI = Mano Izquierda.

Es decir, el conjunto: $J = \{CD, CI, MD, MI\}$ y el torso *T*.

Así también se definen: $\theta = \{\theta_{CD}, \theta_{CI}, \theta_{MD}, \theta_{MI}\}$ y $\varphi = \{\varphi_{CD}, \varphi_{CI}, \varphi_{MD}, \varphi_{MI}\}$.

Con lo anterior, las siguientes ecuaciones permitirán enfrentar el problema de variación en la ubicación del usuario.

$$\sum_{i=1}^{n} D(i) = \sqrt{(J(i)_x - T_x)^2 + (J(i)_y - T_y)^2 + (J(i)_z - T_z)^2}$$

Ecuación 4.1 Distancia entre Puntos del Espacio[32]

$$\sum_{i=1}^{n} \theta(i) = atan2\left(\sqrt{(J(i)_x - T_x)^2 + (J(i)_y - T_y)^2}, (T_z - J(i)_z)\right)$$

Ecuación 4.2 Coordenada Polar θ[33]

$$\sum_{i=1}^{n} \varphi(i) = atan2\big((J(i)_y - T_y), (J(i)_x - T_x)\big)$$

Ecuación 4.3 Coordenada Acimutal φ[34]

Donde *n* es el número de joints considerados (Codos y Manos).

4.4.3.2 Normalización del Tamaño del Usuario

Otro problema en el reconocimiento de gestos deriva de la variación entre las estaturas de los usuarios, es decir, el sistema debe ser capaz de identificar correctamente el gesto realizado independientemente de la mayor o menor distancia entre los joints producto de si el usuario es alto o bajo, respectivamente.

[32] Los puntos del espacio para este trabajo corresponden al Joint (Mano o Codo) y el Torso.

[33] Aplicado para hallar el ángulo formado por el vector (Resultado del Joint en cuestión y Torso) y el eje Z.

[34] Ángulo formado por la proyección en el plano XY del vector (Resultado del Joint en cuestión y el Torso) y el eje X.

Figura 4.21 Variación en tamaños de usuario. Fuente: [58]

Una solución empleada en trabajos de terceros es realizar el entrenamiento varias veces para cada signo, considerando usuarios con diferentes tamaños o características físicas, no obstante, dicha alternativa conlleva mayores costos e incurre en ineficacia e ineficiencia.

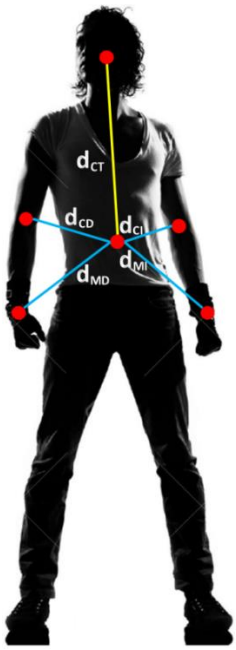


Figura 4.22 Normalización requerida de joints por tamaño del usuario. Fuente: Elaboración propia.

Este trabajo maneja la propuesta ilustrada en la Figura 4.20, la cual consiste en normalizar todas las distancias mediante la Ecuación 4.4, a partir de un factor resultado de la distancia espacial entre los joints de la Cabeza y el Torso (d_{CT}), para hallar este valor se volverá a emplear la Ecuación 4.1.

Entonces se define $D = \{d_{CD}, d_{CI}, d_{MD}, d_{MI}\}$, y la normalización de las distancias se daría por:

$$\sum_{i=1}^{n} D(i) = \frac{D(i)}{d_{CT}}$$

Ecuación 4.4 Normalización de distancias acorde al tamaño de usuario

Donde *n* es la cantidad de distancias a normalizar.

4.4.4 Descriptor de Signo

El Descriptor parametrizará cada gesto para ser almacenado en el Diccionario de Gestos[35] con los demás.

El Descriptor empleado genera tantas filas como frames correspondan al gesto.

Siendo coherente con el resultado de porcentaje de reconocimiento y conclusiones que obtienen en el trabajo de Capilla [31], no se considerará el cálculo y empleo de la Coordenada Polar (θ). La razón se explica a continuación, el objetivo de Capilla [31] es demostrar cuál de sus tres configuraciones se desempeña mejor en reconocimiento de gestos:

1) Coordenadas Cartesianas + Nearest Neighbor-DTW (Cartesian + NN-DTW): El descriptor almacena coordenadas X, Y y Z de los joints de interés, además sólo considera la normalización de la posición del usuario frente al sensor. El clasificador empleado es Nearest Neighbor-DTW.
2) Coordenadas Esféricas + Nearest Neighbor-DTW (Spherical + NN-DTW): El descriptor contiene las coordenadas esféricas de los joints, considera la normalización de la ubicación del usuario. El clasificador usado es Nearest Neighbor-DTW.
3) Coordenadas Esféricas + Nearest Group-DTW (Spherical + NG-DTW): Difiere de la anterior empleando como clasificador al Nearest Group-DTW.

[35] Se denomina así al conjunto de archivos planos que contienen las secuencias de gestos que en algún momento han sido registrados mediante el software.

	Cartesian + NN-DTW			Spherical + NN-DTW			Spherical + NG-DTW		
	H=0.8, E=0.2	H=0.5, E=0.5	H=0.2, E=0.8	H=0.8, E=0.2	H=0.5, E=0.5	H=0.2, E=0.8	H=0.8, E=0.2	H=0.5, E=0.5	H=0.2, E=0.8
x / d	77.381%	80.350%	77.381%	73.810%	78.5714%	77.381%	71.429%	75.000%	72.619%
	77.421%	80.396%	77.381%	73.8095%	78.5714%	77.500%	71.429%	75.000%	72.7381%
y / θ	4.762%	7.143%	8.330%	5.357%	8.928%	10.714%	4.762%	5.952%	8.330%
	4.762%	7.024%	8.214%	5.2381%	9.008%	10.794%	4.643%	6.032%	8.413%
z / φ	71.429%	70.833%	68.452%	94.048%	91.660%	88.690%	91.071%	87.500%	82.143%
	71.429%	70.952%	68.810%	94.405%	92.143%	88.166%	91.4286%	87.980%	82.739%
x,y / d,θ	58.928%	72.619%	75.5952%	57.143%	59.524%	51.191%	64.286%	58.929%	44.643%
	57.857%	72.5794%	75.754%	57.143%	59.524%	51.310%	64.286%	58.929%	44.881%
x,z / d,φ	85.119%	80.357%	74.405%	**95.238%**	93.452%	86.905%	92.262%	91.071%	83.929%
	85.119%	80.357%	74.524%	**95.238%**	93.452%	86.905%	92.262%	91.071%	83.929%
y,z / θ,φ	71.429%	70.833%	69.048%	75.595%	70.238%	60.714%	70.238%	66.670%	54.762%
	71.429%	70.833%	69.405%	75.952%	70.516%	61.071%	70.595%	67.024%	55.952%
x,y,z / d,θ,φ	85.119%	82.738%	75%	**94.643%**	91.660%	80.952%	**94.643%**	91.666%	80.952%
	85.119%	82.738%	75.119%	**94.643%**	91.660%	81.071%	**94.643%**	91.666%	81.071%

Figura 4.23 Porcentaje de reconocimiento según diferentes configuraciones en el trabajo de Capilla. Fuente: [31]

La primera columna indica las características usadas X, Y, Z (Coordenadas Cartesianas) y d, θ, φ (Coordenadas Esféricas), mientras la primera fila referencia el Tipo de Clasificador evaluado. La segunda columna hace alusión a la configuración de pesos de los joints dado: H = Hands (Manos), E = Elbows (Codos). En cada celda existen dos valores, el primero muestra la exactitud de reconocimiento general del sistema, y el segundo la exactitud considerando los resultados individuales de cada signo.

El mejor resultado mostrado en la figura anterior es obtenido considerando las coordenadas cartesianas X y Z, y las coordenadas esféricas *d* y *φ*.

En consecuencia, el descriptor almacenará los valores correspondientes a las coordenadas esféricas *(d y φ)* para cada uno de los joints definidos.

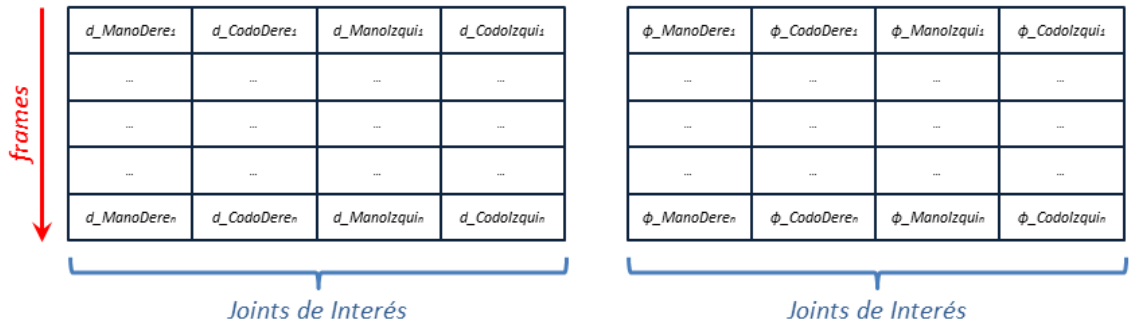


Figura 4.24 Descriptor de signos basado en coordenadas esféricas para cada joint. Fuente: Elaboración propia.

4.4.4.1 Último Frame del Gesto Realizado

Cuando el usuario considere haber terminado un gesto podrá posicionar ambas manos por debajo de la altura correspondiente a su cintura, dado lo anterior, el sistema finalizará el almacenamiento de los frames.

4.4.5 Clasificador

Una vez capturados y parametrizados todos los frames correspondientes al gesto, el sistema identificará si el usuario ha indicado Reconocer Seña o Agregar Seña. En el primer caso, el sistema realiza la clasificación comparándolo con los existentes en el Diccionario de Gestos. Se presenta mayor detalle en 4.4.5.1.

Caso contrario, si el usuario ha indicado Agregar Seña, el sistema almacenará el conjunto de frames ya parametrizados en el Diccionario de Gestos. Mayor detalle se brinda en 4.4.5.2.

4.4.5.1 Reconocer Seña

Según Ibañez [48], otros factores limitantes son la variación de velocidad y destreza de los usuarios al realizar los gestos, sumados a la dificultad para definir reglas en el reconocimiento de gestos complejos. Por ello, en ese mismo trabajo se considera adecuado abordar técnicas de Supervised Machine Learning (Aprendizaje de Máquinas Supervisado) frente a la problemática empleando una base de gestos de entrenamiento para construir un clasificador.

El Clasificador tiene la función de evaluar los movimientos del usuario y determinar su similitud con los gestos entrenados (Almacenados en el Diccionario de Gestos) con la finalidad de entregar el resultado más cercano en lenguaje escrito y hablado. Si bien diversas soluciones ofrecen reconocimientos cercanos al 100% de los gestos evaluados, el desarrollador o investigador se ve obligado a implementar técnicas complejas para incluir el reconocimiento de gestos dentro de su aplicación. [48]

El Clasificador debe lidiar correctamente en la comparación de señas con diferentes cantidades de frames.

Cada vez que la aplicación es iniciada, el sistema cargará los gestos y sus distancias asociadas en el archivo XML (Como se muestra en la Figura 4.25), en caso este fichero

no muestre gestos grabados, el sistema inhabilitará la opción Reconocer Seña hasta que el usuario realice el entrenamiento y almacenamiento de cuando menos un gesto.

Como se definió en capítulos anteriores, el Clasificador se basará en Dynamic Time Warping (DTW), este permite comparar y medir la similitud de dos secuencias que pueden variar en tiempo o velocidad. Valiéndonos además de la experiencia de terceros en tanto a la aplicación de este método en casos de reconocimiento de voz y audios, videos, y aplicaciones gráficas, este método servirá para el reconocimiento de gestos a partir de la data obtenida y procesada en los pasos anteriores.

A continuación, se muestra el pseudocódigo correspondiente DTW.

***Algoritmo* 4. 1**: *Dynamic Time Warping*

```
funcion: DTW (char s[1 ... n], char t[1 ... m])
    entero DTW [0 ... n, 0 ... m]
    entero i, j, cost
    para i desde 1 hasta m hacer
        DTW[0, i] ← ∞
    fin para
    para i desde 1 hasta n hacer
        DTW[i, 0] ← ∞
    fin para
    DTW[0,0] ← 0
    para i desde 1 hasta n hacer
        para j desde 1 hasta m hacer
            cost ← funcionCosto(s[i], t[j])
            DTW[i, j] ← cost +
                    mínimo(DTW[i − 1, j], DTW[i, j − 1], DTW[i − 1, j − 1] )
        fin para
    fin para
    retornar DTW[n, m]
fin funcion
```

Como se puede observar, este algoritmo permite realizar comparaciones sólo para data unidimensional no pudiendo aplicarse directamente al presente trabajo, por lo cual es necesaria la realización de modificaciones para obtener la similitud entre dos secuencias n-dimensionales de frames generados por el Descriptor, donde cada posición a comparar contendrá un arreglo con data similar a:

$$\left[MD_{d_1}, CD_{d_1} MI_{d_1}, CI_{d_1}, MD_{\varphi_1}, CD_{\varphi_1}, MI_{\varphi_1}, CI_{\varphi_1}\right]$$

La modificación consiste en hallar el costo entre ambas secuencias de data. Considerando dos secuencias *S1* y *S2*.

$$cost = \sqrt{\sum_{i=1}^{n} (S1_i - S2_i)^2}$$

Ecuación 4.5 Función costo sin considerar el peso de cada Joint de Interés

Además, es necesario establecer un peso para cada joint, esto permitirá aprovechar mejor la data considerada más importante de los 4 joints definidos (Codos y manos). Estos pesos son aplicados cuando los *costos* son obtenidos de las dos secuencias *S1* y *S2*, entonces la ecuación mostrada sería:

$$cost = \sqrt{\sum_{i=1}^{n} peso_i * (S1_i - S2_i)^2}$$

Ecuación 4.6 Función costo considerando el peso de cada Joint de Interés

Teniendo en cuenta que el peso será el mismo para cada tipo de joint, por lo tanto, no generará inconvenientes. Este valor fluctuará entre 0 y 1, y será definido de la siguiente forma, considerando que el movimiento de los codos es inferior al de las manos y además mayormente paran más pegados (Menor distancia, varían menos) al torso durante la ejecución de los gestos, por lo cual, las manos son las que más describen la realización del gesto. Especificado lo anterior, como se muestra en la Figura 4.23, los mejores resultados se obtendrán cuando los Codos y Manos tengan un peso de 0.2 y 0.8, respectivamente.

Finalmente, los parámetros correspondientes al gesto realizado son comparados con cada una de las secuencias que se han definido como las representativas por cada gesto

almacenado en la etapa de entrenamiento. La evaluación que arroje menor distancia permitirá la clasificación del gesto realizado, posteriormente dicho resultado será mostrado en pantalla de forma escrita y hablada por el software.

4.4.5.2 Agregar Seña

Bajo esta modalidad, el conjunto de frames parametrizados correspondientes a una nueva secuencia de gesto serán almacenados en el Diccionario de Gestos. Cada gesto con un nombre diferente a los existentes será registrado en un archivo plano nuevo, en caso el nombre del gesto a entrenar exista, se procederá a agregarlo al correspondiente archivo, en otras palabras se pueden registrar varias secuencias del mismo gesto, dependiendo de si el usuario desea realizar el entrenamiento valiéndose de diferentes muestras, lo cual resulta importante porque permite contar con variantes del mismo gesto tanto en la forma de ejecutarlo en distintos puntos del área de detección, como con diferentes características físicas al variar de personas gesticuladoras. Contar con variaciones de un gesto permite a las técnicas de reconocimiento considerar diversidad de ejemplos y mejorar su precisión.

Una vez que el usuario ha finalizado la realización del gesto, se agregará un identificador para reconocer correctamente dónde inicia la nueva secuencia en el fichero plano.

Si el gesto es nuevo, se agregará un hijo <Gesto> a la raíz <Gestos> conteniendo como parámetros el Nombre del gesto y el Patrón. Esta información se almacenará en un fichero XML[36] que contendrá la estructura que muestra la siguiente imagen.

[36] Se elige XML (eXtensible Markup Languaje) como formato para representar y estructurar este tipo de datos como consecuencia de ser reutilizable y multiplataforma. Las tareas de ver, agregar y actualizar valores se pueden realizar fácilmente sin la necesidad de implementación de un parser gracias a la existencia de librerías que facilitan ello.

Figura 4.25 Archivo de configuración XML. Fuente: Elaboración propia.

Cada vez que una secuencia es agregada al Diccionario de Gestos se calculará la distancia mediante el Algoritmo 4.1 entre todas las secuencias existentes para el gesto ingresado. La evaluación busca encontrar cuál de todos los entrenamientos realizados para el gesto X presenta menor distancia con el resto, una vez obtenido dicho valor, se actualiza el archivo clasificador XML indicando la posición de la secuencia en su archivo plano y se procede a cargarla en memoria.

Se debe entender por Nombre o Etiqueta del Gesto a cómo este se identifica dentro del Diccionario de Gestos, y cuyo valor será mostrado en pantalla cuando el sistema lo identifique (Clasifique) como tal a un gesto realizado en modo Reconocer Seña.

Las capturas de las interfaces del software propuesto se muestran en el [Anexo 08].

CAPÍTULO V: EVALUACIÓN DEL SISTEMA

Para la realización de pruebas, se deben definir aspectos que permitirán obtener métricas necesarias para determinar si el software ha logrado los objetivos propuestos o en su defecto identificar dónde realizar ajustes y/o cambios para acercarnos y alcanzar las metas deseadas.

5.1 Diseño del Campo de Evaluación

Considerando la información recopilada en 4.2 y 4.3 para lograr verificar que el sistema satisfaga los requerimientos técnicos del usuario, se establece lo siguiente:

1) Personal: Según el esquema de desarrollo mencionado en el Capítulo IV, para la etapa de evaluación existen dos tipos de mediciones que se pueden realizar: Evaluación Objetiva y Subjetiva.

   Para llevar adelante la Evaluación Objetiva se contará con cinco personas de diferentes características físicas y diestras en la realización de los gestos definidos. Con lo anterior se procederá a probar el sistema repetidas veces para obtener información correspondiente a las tasas de reconocimiento y tiempo promedio de procesamiento del sistema en traducir gestos dada cierta cantidad existente en el Diccionario de Gestos. Las evaluaciones se realizarán físicamente en el área de atención al público de la Institución Educativa N° 20915 – Pucará, entidad en la cual se realizaron las entrevistas que permitieron establecer los términos a traducir.

   La Evaluación Subjetiva requiere participación de personas no oyentes que conozcan y empleen lengua de señas para comunicarse, quienes además no deben haber sido involucradas en la Evaluación Objetiva. Así también, la participación de personas oyentes representando el rol de usuario del sistema. Para este tipo de evaluación se tiene en cuenta la información proveniente de los formularios o encuestas llenados con opiniones subjetivas de los implicados.

   Sin embargo, acorde al alcance definido y las etapas establecidas como trabajo futuro, se limita la evaluación al aspecto Objetivo.

   Mayor información de ambos tipos de evaluaciones se ofrece en 5.2 y 5.3.

2) Gestos: Las señas a reconocer bajo el alcance definido se detallan en el punto 4.3.2. Estos corresponden a palabras o expresiones individuales debiendo ser reconocidos de forma independiente fuera de un contexto físico determinado.

5.2 Evaluación Objetiva

La evaluación debe incluir métricas objetivas del sistema propuesto. En esta sección se medirá la exactitud del sistema para la técnica de reconocimiento implementada y configuración de parámetros otorgado.

Con la intención de evaluar la tasa de reconocimiento obtenida dado un mínimo entrenamiento y realizar su comparación frente a propuestas de terceros, se opta por llevar a cabo el entrenamiento con una sola persona no oyente diestra en lengua de señas quien realizará cinco veces cada gesto frente al sensor. Se establece dicha cantidad siguiendo los trabajos mencionados en el Estado del Arte.

Considerando la información plasmada en el punto 5.1 acerca de la cantidad de personal y número de gestos definidos para la etapa de pruebas, y siguiendo las investigaciones citadas en el Estado del Arte, se establece determinar la tasa de reconocimiento con cinco personas de diferentes características físicas quienes deben contar con destreza en los gestos definidos, realizándolos cinco veces cada uno.

5.2.1 Entrenamiento del Software

Para el correcto funcionamiento del software y su evaluación, se procederá a entrenar al sistema con los gestos establecidos.

A nivel de implementación software, cada tipo de seña contará con su propio archivo de almacenamiento conteniendo todas las secuencias correspondientes asociadas a cada entrenamiento. Cada vez que se realice el entrenamiento de una nueva muestra de un gesto existente, esta se almacenará y comparará con las secuencias almacenadas mediante el algoritmo DTW. La comparación se realiza con la finalidad de determinar cuál de todas presenta menor distancia con el resto para ser considerada como la secuencia modelo o mejor representación del gesto.

La persona no oyente definida para esta actividad, quien habitualmente emplea Lengua de Señas Peruana para comunicarse, realizará cada gesto cinco veces desde la ubicación central del área de detección del sensor, llevándolas a cabo con variaciones en destreza y

velocidad, las cuales son características que generalmente se agregan de forma subconsciente cuando cualquier persona emplea lengua de señas.

Después del análisis y evaluaciones realizadas en el Estado del Arte, se concluyó elegir DTW como método que permite facilitar el uso del software traductor sin requerir un entrenamiento exhaustivo en caso se deseen agregar más gestos al diccionario y permitiendo sumar la cantidad deseada sin reducir la performance, ni requerir ajustes técnicos, lo cual permitiría al software contar con los gestos necesarios para ser empleado en cualquier contexto. En general, los trabajos de terceros han empleado métodos estadísticos (Hidden Markov Models, Redes Neuronales Artificiales, entre otros) que conllevan mayores costos de programación y recursos en general para obtener resultados, y además exigen a los usuarios realizar mayores iteraciones de entrenamiento para permitir que los gestos sean correctamente reconocidos.

5.2.2 Costo Económico

Para llevar a cabo el desarrollo del software, se decidió recurrir convenientemente al uso de herramientas tecnológicas de libre acceso con las intenciones de reducir los costos y volver al proyecto factible y viable, así mismo convertirlo en una opción accesible económicamente en contraste con propuestas de terceros.

Se emplearon las siguientes herramientas:

1) Visual Studio Community, empleado como IDE de desarrollo. Se puede descargar de forma gratuita del portal web de Microsoft.
2) SDK, driver que permite la conexión y reconocimiento entre el SO y el sensor. Se puede descargar gratis desde el portal web del sensor.
3) Sensor Kinect, hardware adquirido a 50 dólares, el cual permite realizar el seguimiento de los movimientos humanos.

En total, se considera un costo monetario de 50 dólares, un gasto único e ínfimo en contraste al cobro recurrente que se daría con la contratación de intérpretes.

5.2.3 Tasa de Reconocimiento

Para obtener esta medición se recurrirá al apoyo de cinco personas diestras en la realización de los gestos definidos, quienes ejecutarán la aplicación bajo la opción

Reconocer Seña. Dichas personas han sido instruidas por el no oyente encargado del entrenamiento del software.

Cada gesto será efectuado cinco veces por cada usuario, quien las llevará a cabo desde ubicaciones aleatorias en el área de detección. Esto se da con la intención de comprobar la robustez del sistema frente a variaciones en los parámetros de entrada dependientes de las características físicas y posición del usuario.

La tasa de reconocimiento será obtenida mediante la siguiente ecuación:

$$Tasa\ de\ Reconocimiento_{Gesto} = \frac{Positivos_{Gesto}}{Positivos_{Gesto} + Negativos_{Gesto}} * 100$$

Ecuación 5.1 Ecuación para hallar la Tasa de Reconocimiento de un Gesto

Esta ecuación busca medir la exactitud de reconocimiento de cada gesto, donde:

- $Positivos_{Gesto}$: Representa la cantidad de pruebas cuyo resultado ha sido correcto.
- $Negativos_{Gesto}$: Representa la cantidad de pruebas cuyo resultado ha sido incorrecto.

Para llevar adelante esta etapa, se consideró dividirlo en dos fases:

1) Evaluación Preliminar

   Realizada con la intención de identificar posibles problemas y limitaciones del software frente al reconocimiento de gestos en diferentes contextos, para posteriormente determinar e implementar las soluciones pertinentes.

   Esta evaluación involucró la participación de dos personas voluntarias quienes fueron adiestradas por el encargado del entrenamiento del sistema.

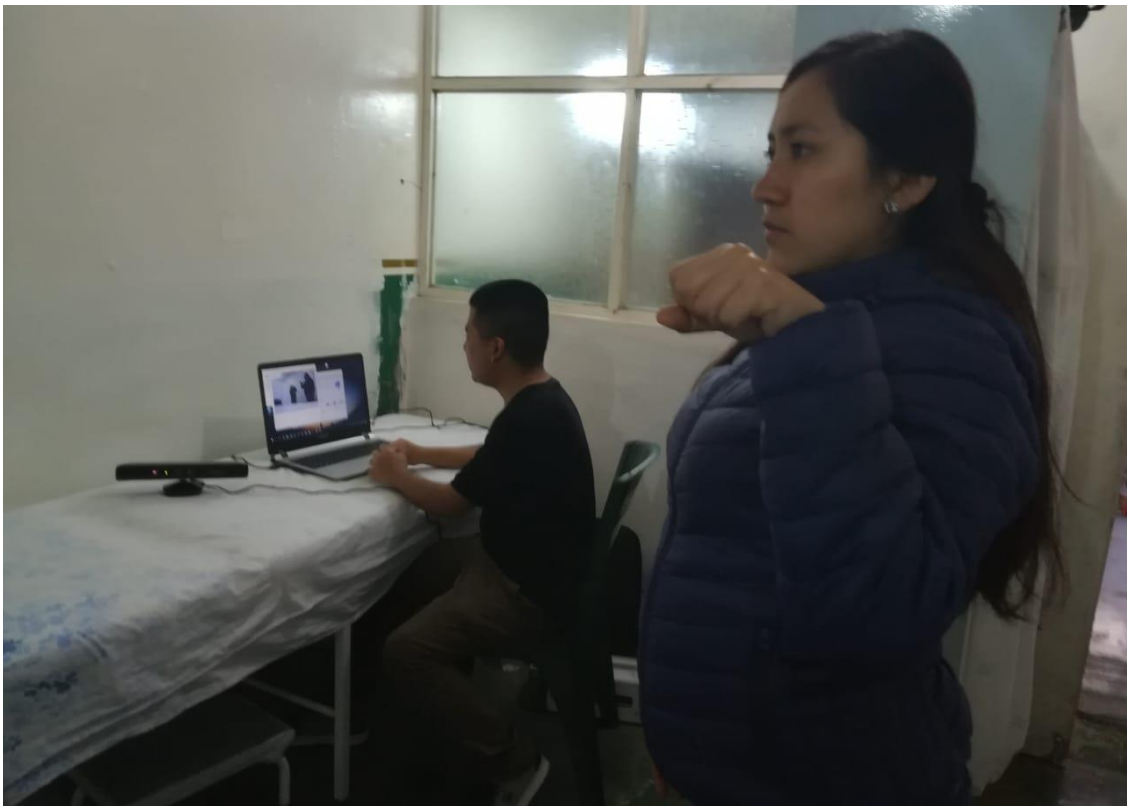


Figura 5.1 Escenario empleado para la Evaluación Preliminar 01. Fuente: Elaboración propia.

Con la intención de probar la robustez de la propuesta software frente a variaciones de iluminación y contexto, se vio por conveniente realizarla en lugares aleatorios relacionados con la atención del público.

Durante esta evaluación y tal como se observa en las imágenes 5.2 y 5.3, la actividad consistió en la realización de los gestos definidos en los capítulos anteriores, de tal forma estos se llevasen a cabo 5 veces por cada persona.

Figura 5.2 Evaluación Preliminar 01 de la propuesta software. Fuente: Elaboración propia.

Como se presentan en las imágenes 5.2 y 5.3, se detectó un inconveniente que originaba una menor tasa de reconocimiento. El problema es resultante de la oclusión entre distintas partes del cuerpo que puede darse durante la realización de cualquier gesto. Por ejemplo, en las figuras mencionadas se detectan a simple vista la presencia eventual de unos puntos amarillos los cuales representan las posiciones de las manos y brazos. Durante el análisis de las posibles causantes se encontró que aparecen cuando partes del mismo cuerpo quedan ocultas ante el sensor o sus posiciones se cruzan con la forma del esqueleto dibujado por el Skeleton Tracking. Esto último se tiene en la siguiente imagen, donde el colaborador posiciona su brazo y mano izquierda a la altura de su hombro derecho.

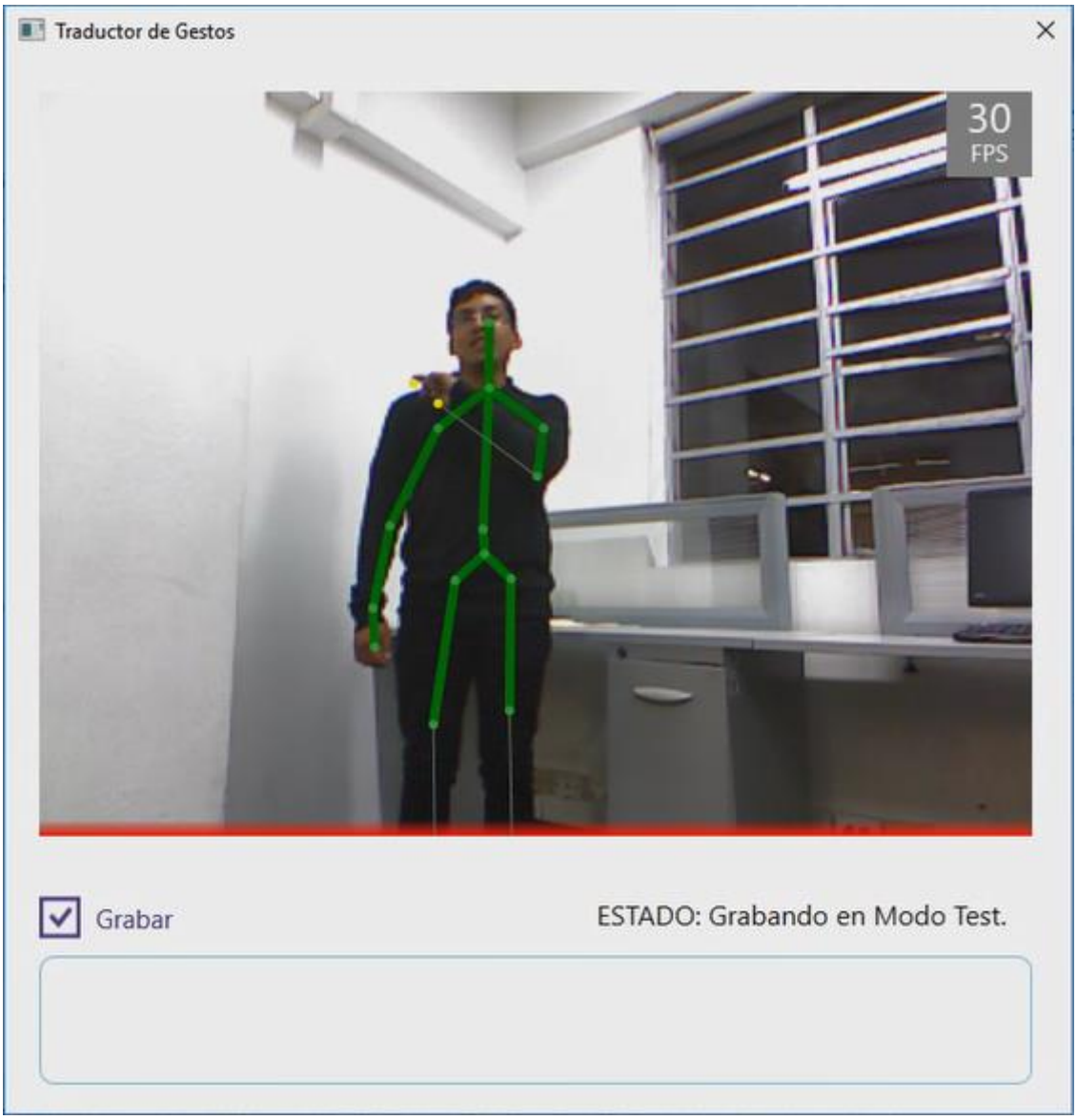


Figura 5.3 Evaluación Preliminar 02 de la propuesta software. Fuente: Elaboración propia.

Dicho inconveniente originaba la captura y almacenamiento incorrecto de datos generados durante el seguimiento de los gestos. Los datos provenientes de las partes representadas con puntos amarillos eran almacenados con valor cero por el Descriptor, lo cual desembocaba en malas clasificaciones realizadas por DTW cuando se intentaba reconocer gestos con el aplicativo.

Para solucionar el inconveniente se realizó el análisis del código implicado en la detección de las partes del cuerpo humano y formación del Skeleton Tracking, finalmente se solucionó el inconveniente modificando una condicional que limitaba la captura de datos cuando se producía oclusión y generaban los puntos amarillos.

2) Evaluación

Finalizada la Evaluación Preliminar, se procedió con la Evaluación del aplicativo. Con intenciones de evaluar la tasa de reconocimiento del software en el contexto de aplicación al cual ha sido enfocado, es decir, áreas de atención al público, especialmente el ambiente de recepción de colegios, se acordó y obtuvo la colaboración de personal de la Institución Educativa N° 20915 – Pucará, ubicado en Cucuya, distrito de Santo Domingo de Olleros, provincia de Huarochirí.

Figura 5.4 Área de recepción de la Institución Educativa N° 20915. Fuente Elaboración propia.

La evaluación se realizó mediante la colaboración de 5 personas quienes previamente fueron adiestradas en Lengua de Señas Peruana por la persona encargada del entrenamiento para la realización de los gestos definidos en este trabajo. Dichas personas cuentan con diferentes características físicas y además se les solicitó la ejecución de los gestos considerando diferentes velocidades.

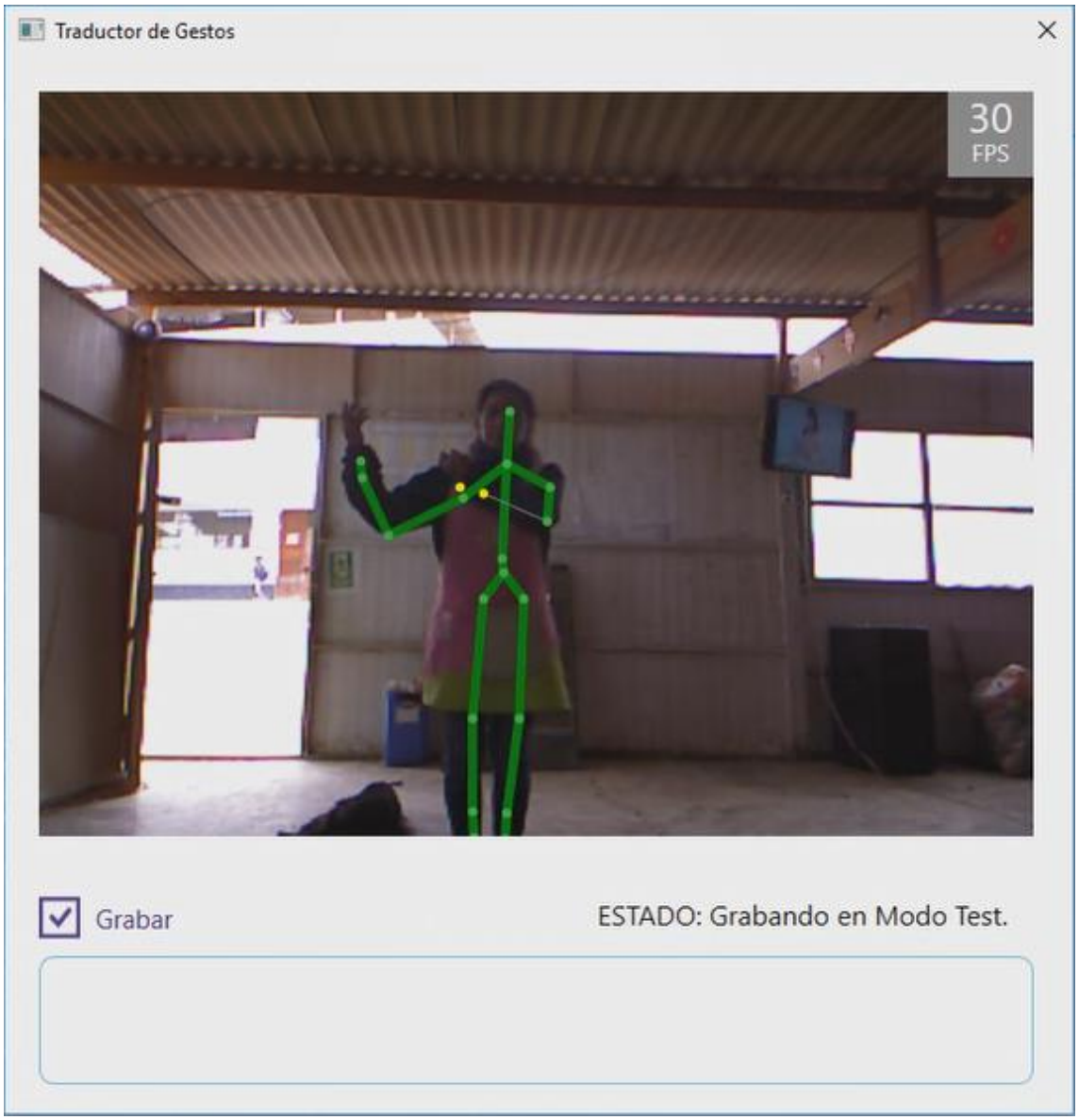


Figura 5.5 Evaluación 01 de la propuesta software. Realización del gesto “Espere”. Fuente: Elaboración propia.

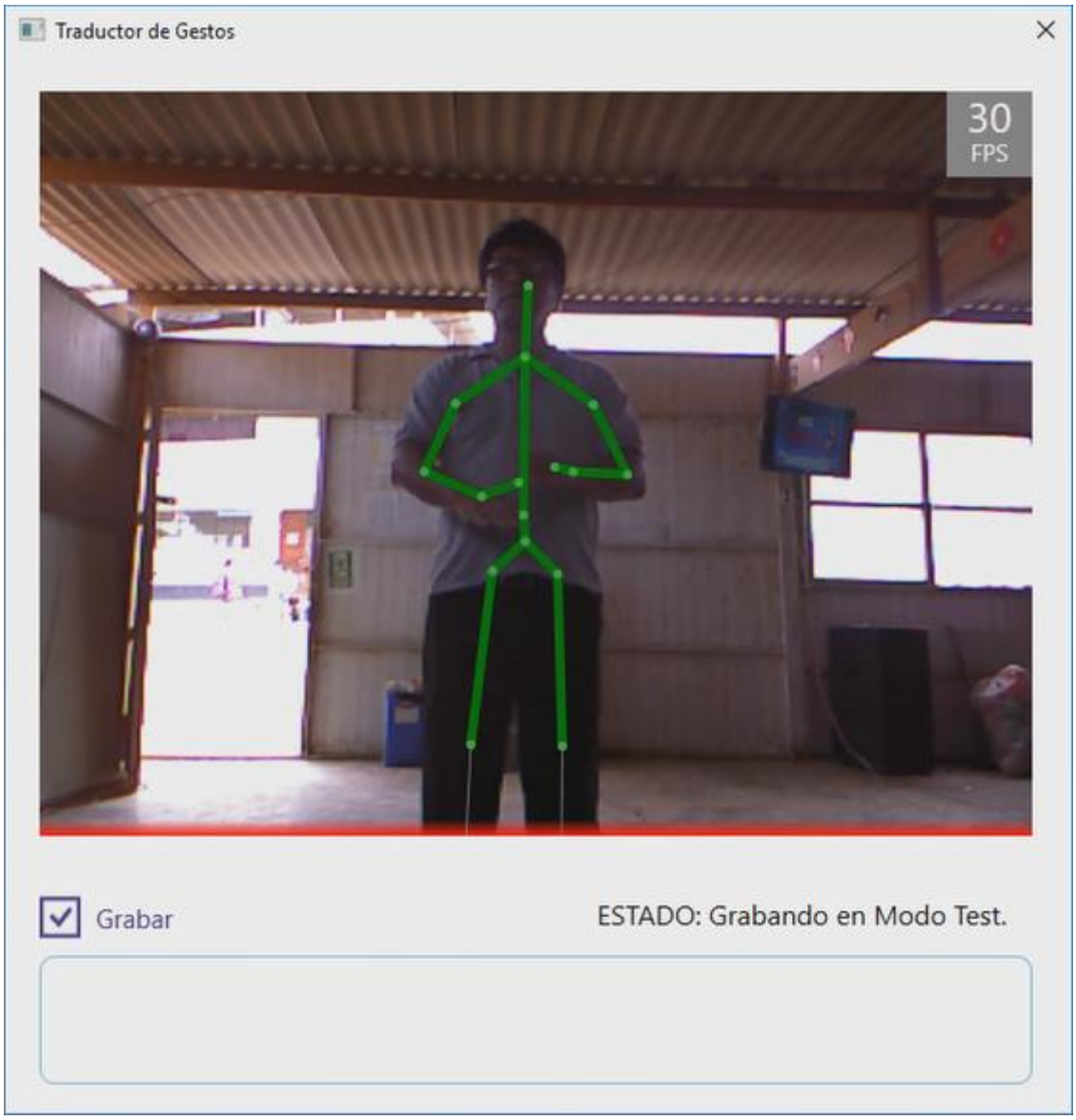


Figura 5.6 Evaluación 02 de la propuesta software. Realización del gesto “Terminado”. Fuente: Elaboración propia.

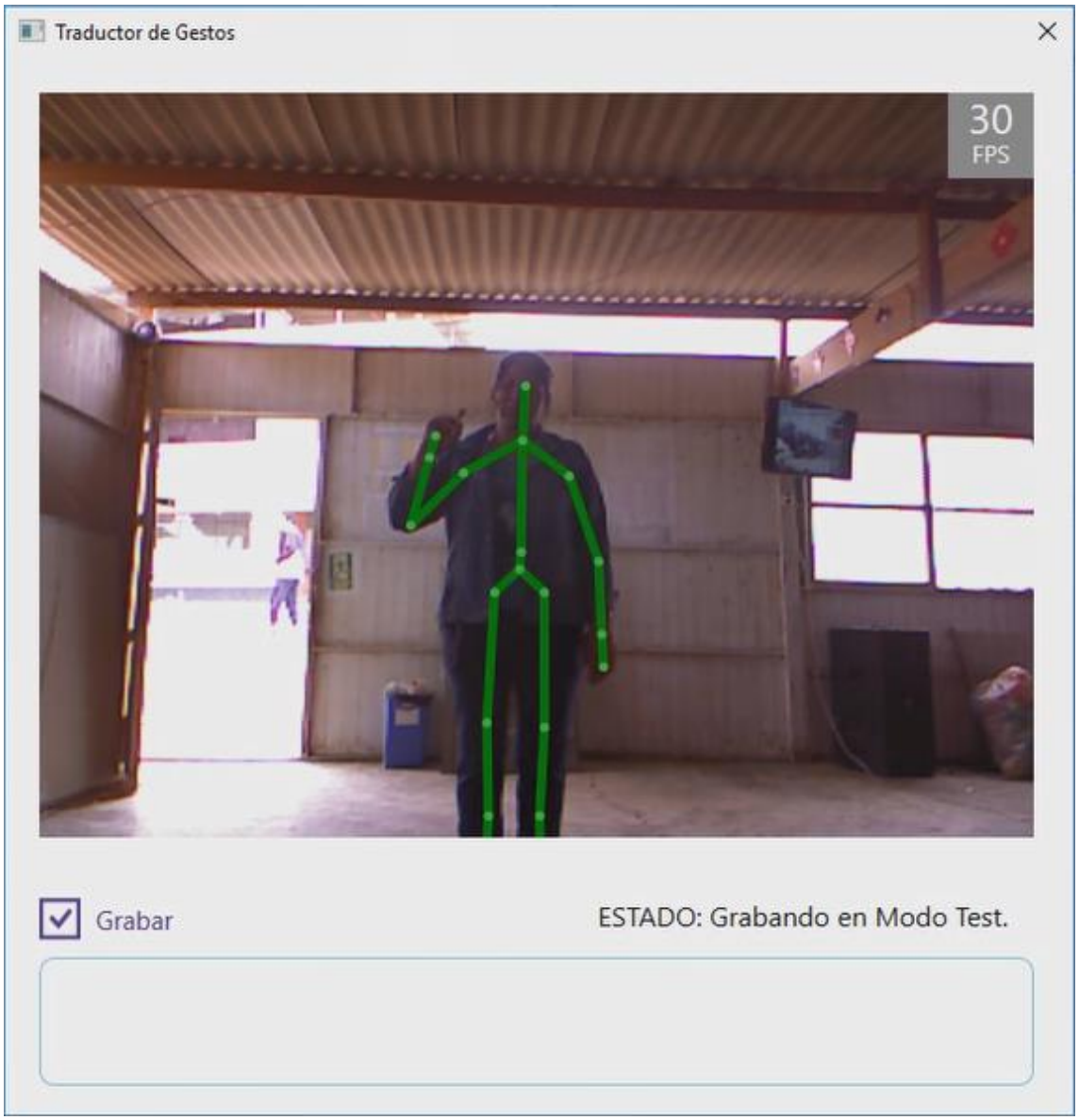


Figura 5.7 Evaluación 03 de la propuesta software Realización del gesto “Docente”. Fuente: Elaboración propia.

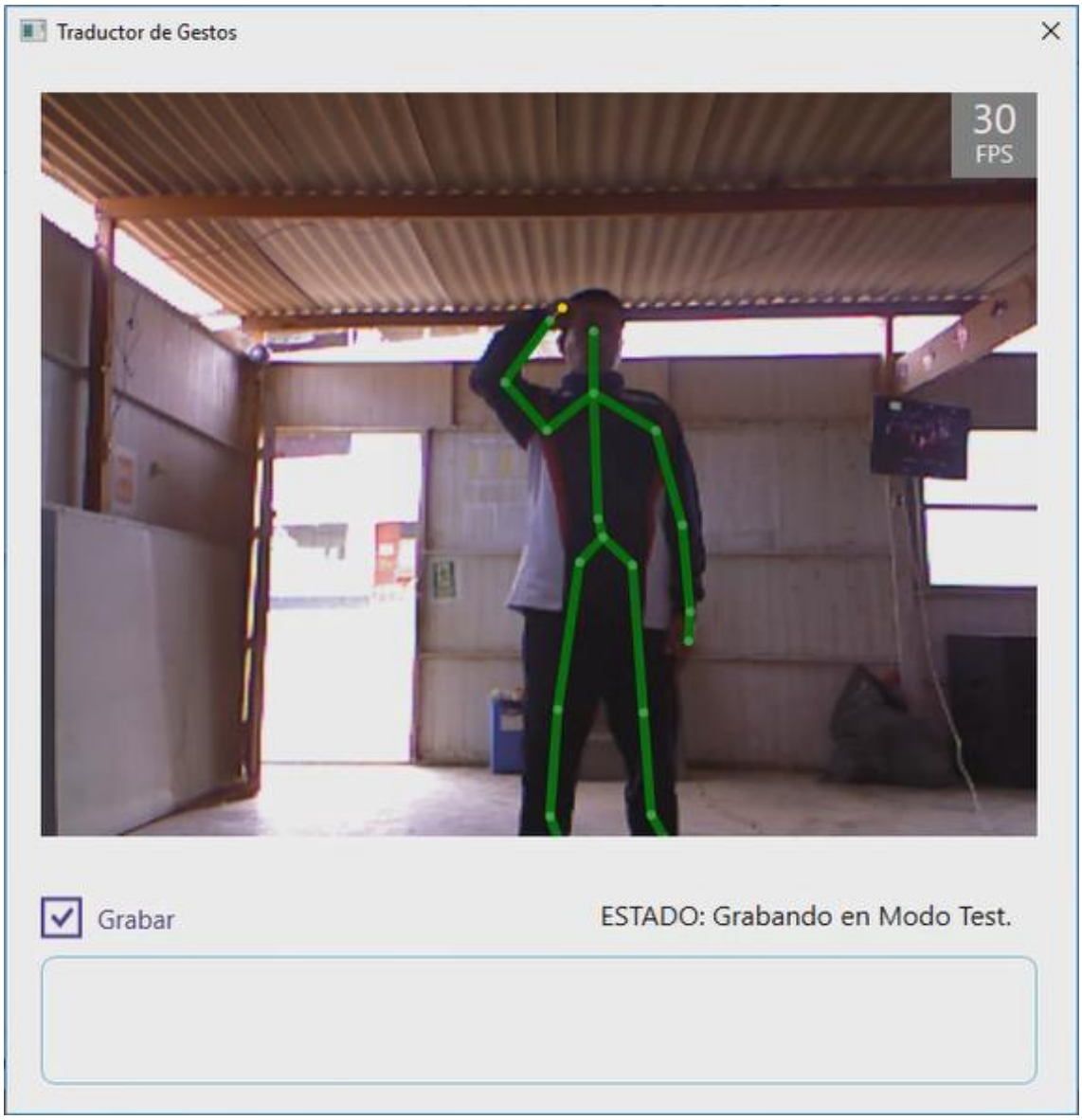


Figura 5.8 Evaluación 04 de la propuesta software. Realización del gesto “Indisciplina”. Fuente: Elaboración propia.

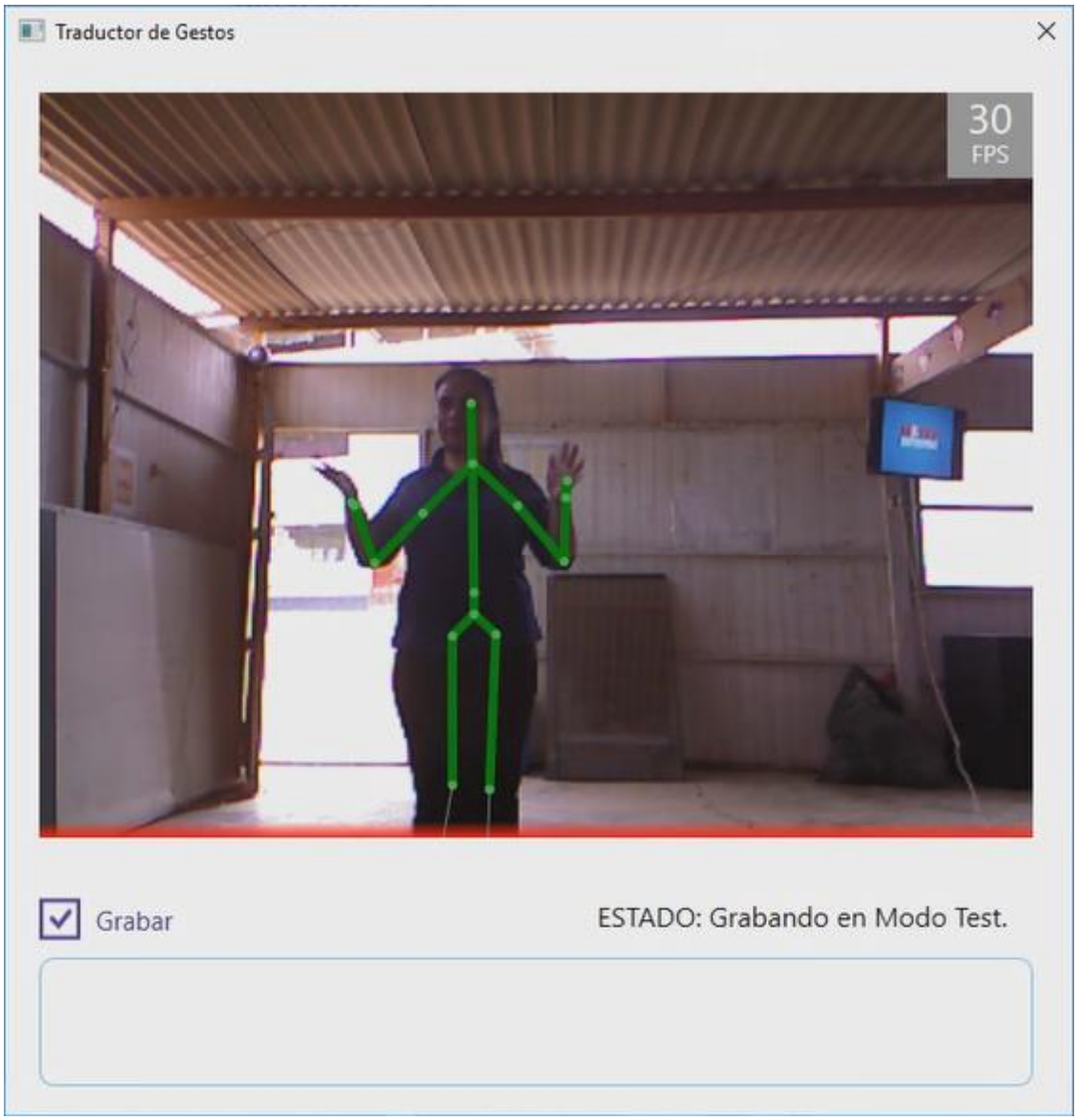


Figura 5.9 Evaluación 05 de la propuesta software. Realización del gesto “No se encuentra”. Fuente: Elaboración propia.

Como se mencionó anteriormente, cada usuario realizará 5 veces los gestos definidos, es decir por cada gesto habrá 25 pruebas en total.

La siguiente tabla muestra el detalle de clasificación de los gestos durante la evaluación, permitiendo conocer en qué cantidad los gestos fueron reconocidos correctamente, y en caso contrario, cuál fue el nombre asignado al gesto de forma errónea.

	Hola	**Bienvenido**	**Director**	**Curso**	**Docente**	**Siéntese**	**Espere**	**Indisciplina**	**¿Qué grado?**	**Terminado**	**No se encuentra**
Hola	**25**										
Bienvenido		**25**									
Director			**25**								
Curso				**25**							
Docente					**25**						
Siéntese						**25**					
Espere					1		**24**				
Indisciplina								**25**			
¿Qué grado?			1					1	**23**		
Terminado				1						**23**	1
No se encuentra											**25**

Tabla 5.1 Errores en clasificación del software. Fuente: Elaboración propia.

Del mismo modo, a continuación, se presenta la tabla resumen de las tasas de reconocimiento obtenidas en la ejecución de las pruebas:

Seña	**Resultado Prueba**		**Tasa de Reconocimiento (%)**
	Positivos	**Negativos**	
Hola	25	0	**100.00**
Bienvenido	25	0	**100.00**
Director	25	0	**100.00**
Curso	25	0	**100.00**
Docente	25	0	**100.00**
Siéntese	25	0	**100.00**
Espere	24	1	**96.00**
Indisciplina	25	0	**100.00**
¿Qué grado?	23	2	**92.00**
Terminado	23	2	**92.00**
No se encuentra	25	0	**100.00**
		Total	**98.18**

Tabla 5.2 Tasa resumen de reconocimiento de los gestos definidos. Fuente: Elaboración propia.

La primera columna de la Tabla 5.2 detalla los gestos definidos para la evaluación. La segunda columna se encuentra subdividida mostrando el número de pruebas correcta (Positivos) e incorrectamente (Negativos) reconocidas. Mientras la última columna indica

la tasa de reconocimiento de cada gesto expresado en porcentaje, esta medición resulta de los datos plasmados en segunda columna y aplicación de la Ecuación 5.1. Para conocer si se ha satisfecho los requerimientos técnicos establecidos puntos anteriores se agregó una última fila encabezada por la denominación "Total" la cual resulta del promedio aritmético de los porcentajes obtenidos en la última columna.

Como se detalla en la Tabla 5.2, la tasa de reconocimiento de cada gesto varía en el rango de 92 a 100%.

Aquellos que han presentado menor tasa de reconocimiento son tres ("Espere" con 96%, "¿Qué grado?" y "Terminado" con 92% cada uno), los cuales han resultado de una incorrecta captura de datos durante la ejecución de gestos, lo anterior sumado a su relativa semejanza con otros gestos han conllevado a obtener clasificaciones erróneas.

El algoritmo DTW es empleado para comparar dos secuencias de gestos sin importar el tamaño (Cantidad de imágenes o frames) con las que cuente cada uno. Entonces, la propuesta software cuenta con la posibilidad de medir la diferencia entre dos gestos realizados, sin embargo, el algoritmo puede generar clasificaciones erróneas, convirtiendo en necesaria una correcta configuración del Descriptor, en ese aspecto, como se demostró en el Capítulo IV, se ha optado por seguir y aplicar las conclusiones demostradas en [31].

Los datos almacenados en el Diccionario de Gestos generados por nuestro software en la Etapa de Entrenamiento y los pertenecientes a otros trabajos como los citados en el Estado del Arte cuentan con diferencias de toda clase. Es decir, pueden emplear data provista por diferentes tipos de inputs (Kinect en nuestro caso, cámaras de diferentes tipos y guantes de datos en otros) y una configuración otorgada a la parametrización de los datos (En nuestro caso realizado por el Descriptor), razones por las cuales los diccionarios definidos en otros trabajos son heterogéneos, al igual que la cantidad y el listado de señas empleados. Lo anterior dificulta e imposibilita emplear sus datos para realizar comparaciones directas sobre los resultados. No obstante, el objetivo principal de esta tesis no es realizar comparaciones con respecto al Estado del Arte, sin embargo, dado que las Tasas de Reconocimiento son la misión principal en trabajos de terceros, se considera interesante mencionar las ventajas y desventajas halladas al realizar ciertas equiparaciones con el enfoque propuesto en esta tesis, ello se detalla en el punto 5.2.4.

5.2.4 Tiempo de Procesamiento

Este trabajo plantea realizar la medición del tiempo de procesamiento de cada gesto cuando el software se encuentra en modo Reconocimiento, considerando la cantidad de gestos almacenados.

Los gestos con los cuales se ha trabajado son sólo una cantidad representativa de la totalidad de términos y expresiones que puede llegar a comprender una Lengua de Señas completamente parametrizada y almacenada en un Diccionario de Gestos. Sin embargo, con la intención de conocer si la propuesta software es capaz de satisfacer los requerimientos del cliente en torno al procesamiento y reconocimiento de gestos en tiempo real, se considera conveniente evaluar y estimar los tiempos de respuesta del software para mayores cantidades de gestos almacenados en el Diccionario de Gestos. Es decir, se evaluará cuánto tarda el sistema cuando se tiene uno, dos, tres hasta diez gestos registrados en el diccionario, posteriormente, se buscará encontrar una regla o función de estimación para conocer los tiempos aproximados de procesamiento cuando se tenga mayor número de gestos almacenados. Del resultado obtenido se procederá a determinar si en etapas futuras es necesaria la optimización de los métodos, pasos realizados y/o código implementado al ejecutarse el Modo Reconocimiento.

En total serán diez mediciones de tiempo por cada gesto, realizados por los cinco usuarios especificados en el ítem anterior, cada usuario estará a cargo de dos pruebas por gesto.

A continuación, se muestra la tabla resumen de las mediciones obtenidas en esta evaluación:

N° de Gestos Almacenados	**Tiempo Promedio de Reconocimiento (Milisegundos)**
1	**19.0011**
2	**39.0022**
3	**47.0027**
4	**74.0042**
5	**98.0031**
6	**108.0063**
7	**119.0068**
8	**138.0079**
9	**148.0085**
10	**181.0103**

Tabla 5.3 Tiempo promedio de reconocimiento según cantidad de gestos. Fuente: Elaboración propia.

En los resultados se observa que el tiempo de reconocimiento tiende a aumentar de forma proporcional al número de gestos almacenados.

Una vez que el usuario ingresó un gesto en Modo Reconocimiento y ha sido parametrizado por el sistema, se procede a compararlo con las secuencias representativas de cada gesto almacenadas en memoria, estas secuencias han sido obtenidas de los ficheros de almacenamiento previamente leídos. Cabe resaltar que la comparación se realiza mediante el algoritmo DTW.

Para hallar el tiempo promedio necesario en el reconocimiento de gestos cuando existan N gestos almacenados en el diccionario, se ha procedido a distribuir los datos de la Tabla 5.3 en la siguiente gráfica.

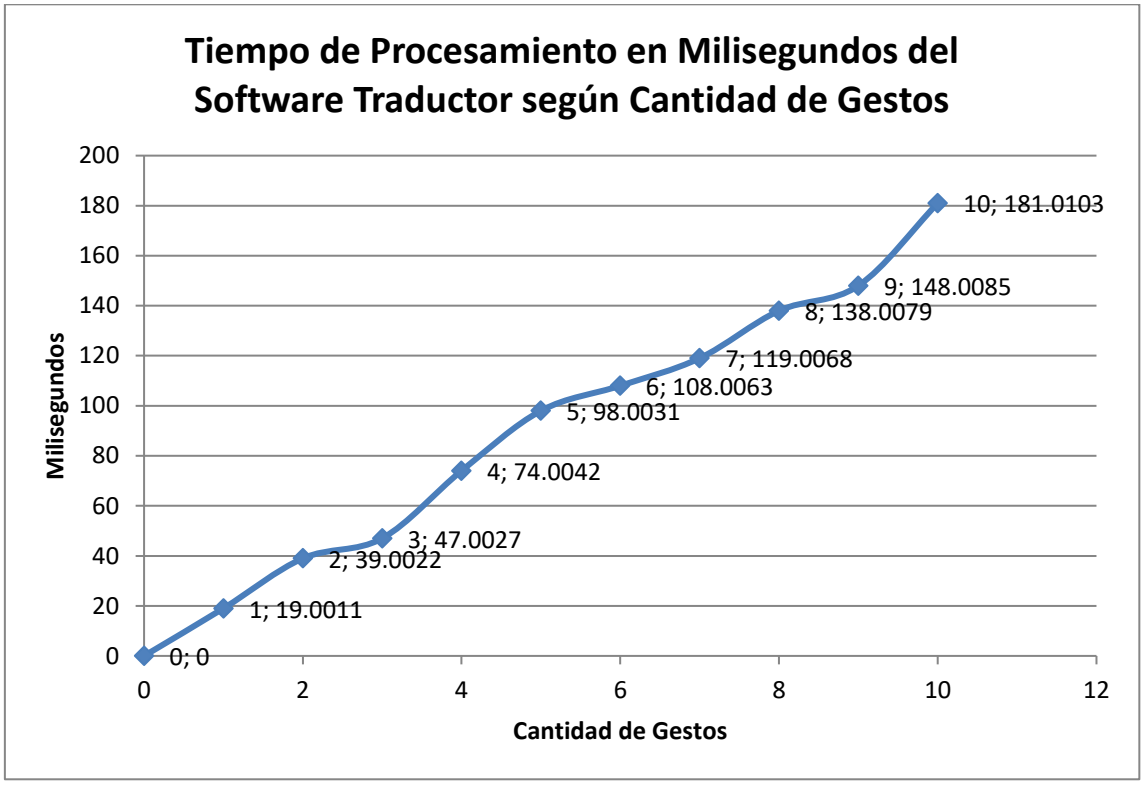


Figura 5.10 Tiempo de procesamiento en milisegundos del software según cantidad de gestos. Fuente: Elaboración propia.

Un sistema orientado al usuario cuenta con tres segundos como máximo para realizar el procesamiento de información y ofrecer el resultado al usuario, caso contrario éste percibirá la demora, perderá el interés y atención, no satisfaciendo el requerimiento de realizar las traducciones en tiempo real. Lo anterior es un escenario no aceptado y para determinar la cantidad de gestos los cuales estando en el Diccionario de Gestos conducirían al sistema a entregar la respuesta en un tiempo cercano a los tres segundos,

se procederá a aplicar el Método de Regresión Lineal por Mínimos Cuadrados al observarse que existe una tendencia lineal en la distribución de los puntos en la gráfica mostrada en la figura anterior.

Se conoce que la relación buscada tiene la forma de una recta, cuya ecuación es:

$$y = mx + b$$

Ecuación 5.2 Ecuación de la recta

En donde las constantes a determinar son: ***m*** la pendiente de la recta y ***b*** la ordenada en el origen (Intercepto). Y ***x*** es el número de gestos, ***y*** es la cantidad de milisegundos.

Primero construimos la tabla de la siguiente forma:

	x_i	y_i	$x_i y_i$	x_i^2	y_i^2
	0	0	0	0	0
	1	19.0011	19.0011	1	361.0418
	2	39.0022	78.0044	4	1521.1716
	3	47.0027	141.0081	9	2209.2538
	4	74.0042	296.0168	16	5476.6216
	5	98.0031	490.0155	25	9604.6076
	6	108.0063	648.0378	36	11665.3608
	7	119.0068	833.0476	49	14162.6184
	8	138.0079	1104.0632	64	19046.1805
	9	148.0085	1332.0765	81	21906.5161
	10	181.0103	1810.1030	100	32764.7287
Σ	**55**	**971.0531**	**6751.3740**	**385**	**118718.1010**

Tabla 5.4 Aplicación del Método de Regresión Lineal por Mínimos Cuadrados. Fuente: Elaboración propia.

Calculamos la Pendiente e Intercepto:

$$m = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}, b = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

Donde *n* es el número de mediciones a considerar, en este caso son 11. Entonces, se tiene:

$$m = \frac{(11)(6751.3740) - (55)(971.0531)}{(11)(385) - (55)^2} = \frac{74265.114 - 53407.9205}{4235 - 3025}$$

$$= \frac{20857.1935}{1210} = 17.23735$$

$$b = \frac{(385)(971.0531) - (55)(6751.3740)}{(11)(385) - (55)^2} = \frac{373855.4435 - 371325.57}{4235 - 3025}$$

$$= \frac{2529.8735}{1210} = 2.0908$$

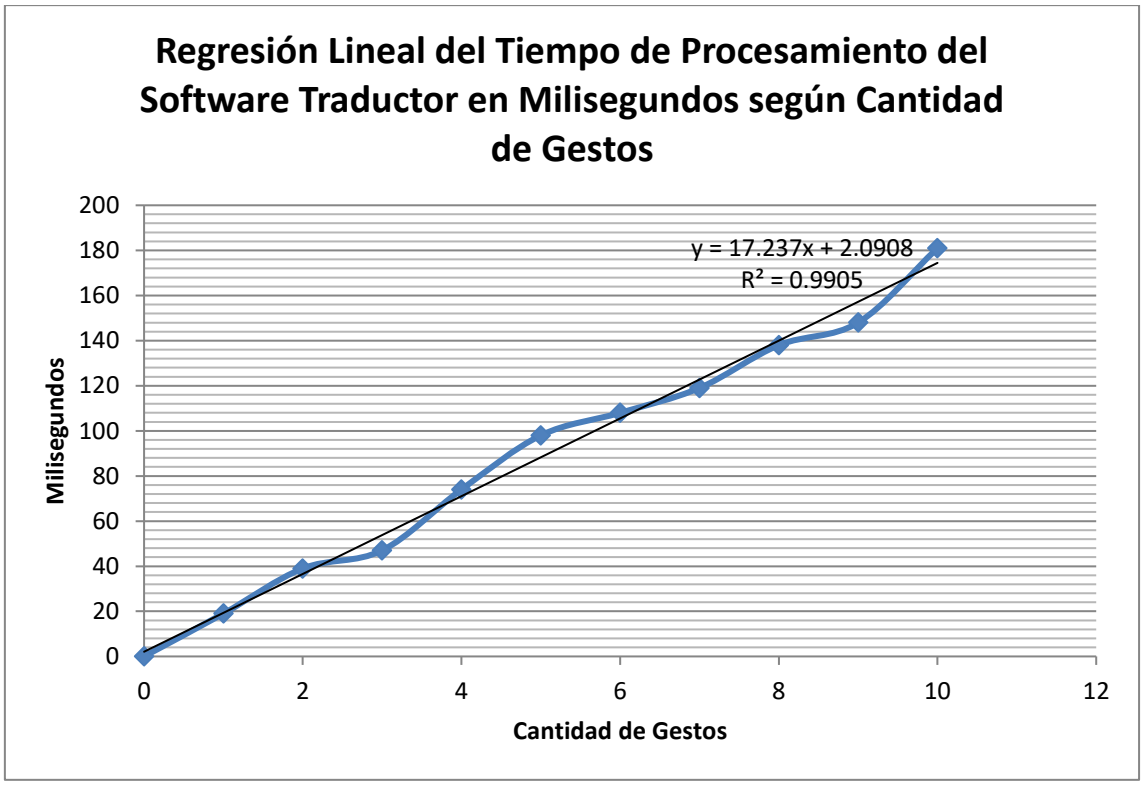


Figura 5.11 Regresión Lineal del Tiempo de Procesamiento del Software Traductor en Milisegundos según Cantidad de Gestos. Fuente: Elaboración propia.

Además, hallaremos el *Coeficiente de Correlación*:

$$r = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{(n\sum x_i^2 - (\sum x_i)^2)(n\sum y_i^2 - (\sum y_i)^2)}}$$

$$r = \frac{20857.1935}{\sqrt{(1210)\big((11)(118718.1010) - (971.0531)^2\big)}} = \frac{20857.1935}{\sqrt{(1210)(362954.987)}}$$

$$= \frac{20857.1935}{20956.5153} = 0.99526 \rightarrow r^2 = 0.9905$$

Cuando más cercano a 1 sea el valor de r^2, mejor es el ajuste. Dado lo anterior y el valor de 0.9905, se verifica que la regla $y = 17.23735x + 2.0908$ nos permitirá estimar cuántos gestos deben estar almacenados en el sistema para ser procesados en tres segundos ante una petición de reconocimiento de gesto en el Modo Reconocimiento.

Entonces:

$$3000 = 17.23735x + 2.0908 \rightarrow 2997.9092 = 17.23735x \rightarrow x = \frac{2997.9092}{17.23735}$$

$$x = 173.9194 \rightarrow x \cong 174\ gestos$$

Es decir, aproximadamente con 174 secuencias de gestos cargados en memoria, el sistema entregaría respuestas en un tiempo estimado de tres segundos. Aparentemente es una vasta cantidad de gestos, no obstante, el replanteamiento de la forma en que se realiza la clasificación y optimización del código se considera conveniente y se dará en un trabajo futuro.

5.3 Análisis y Presentación de Resultados

5.3.1 Análisis de Resultados

Las pruebas objetivas en este trabajo se realizaron sobre los gestos definidos obteniendo data desde la posición y desplazamiento en tiempo real de los Joints de Interés. El sistema no está habilitado para trabajar con información proveniente de la postura de las manos y expresiones faciales, como consecuencia si dos gestos realizan la misma trayectoria el sistema tendrá inconvenientes para discernir correctamente cuál es el gesto realizado. Cabe resaltar que las características de las cuales carece el software serán agregadas en siguientes etapas tal como se menciona en el apartado Trabajos Futuros.

De los trabajos revisados en el Estado del Arte, el más cercano es Capilla [31] quien creó y empleó 14 gestos diferentes que fueron almacenadas en su Diccionario de Gestos. Para el entrenamiento realizó en total 70 muestras (Cinco por gesto), llevadas a cabo en la misma ubicación por cuatro personas de diferentes estaturas. El objetivo del trabajo fue comparar distintas configuraciones del algoritmo DTW, de forma conjunta con los pesos otorgados a los joints de interés, para fundamentar y demostrar cuál de ellas obtiene una mejor performance. Los resultados son mostrados en la Figura 4.23 donde se observa que la mejor configuración logra un porcentaje de reconocimiento del 95.238% (Promedio aritmético obtenido para todos los gestos trabajados).

El promedio aritmético de las tasas de reconocimiento de este trabajo es 98.18% de eficacia en clasificación de las secuencias de gestos. Este resultado es ligeramente superior al mostrado en Capilla [31] siendo consecuencia directa de la diferencia entre las trayectorias de los gestos, y la parametrización otorgada a los datos capturados, los cuales posteriormente son almacenados en caso se encuentre en Modo Entrenamiento. Una consideración especial es que el trabajo de Capilla creó y seleccionó gestos ajenos a una

lengua de señas en particular, no guardando similitud entre sí y evitando conscientemente los problemas de entornos reales mencionados en este trabajo.

Esta tesis ha trabajado con gestos pertenecientes a la Lengua de Señas Peruana empleando convenientemente términos comúnmente usados en la interacción entre personas en las áreas de atención y/o recepción del público, específicamente en ambientes de instituciones educativas.

Comentado lo anterior y bajo el alcance definido, se recalca la importancia de continuar el proyecto acorde a lo especificado en Trabajos Futuros con el fin de lograr un Intérprete de Lengua de Señas, considerando las posturas de manos y torso, expresiones faciales, desplazamiento de brazos y manos, así como la sintaxis propia con la que cuenta toda lengua de señas.

Mientras, en el caso del Tiempo de Reconocimiento de Gestos, se estima que el software dará las respuestas en un tiempo aceptable de máximo 3 segundos hasta con 174 gestos almacenados. En primera instancia puede aparentar ser adecuado, pero teniendo en mente futuras etapas del software se considera necesaria la readecuación de la forma cómo están implementados los componentes software.

5.3.2 Presentación de Resultados

A continuación, se muestran los objetivos específicos de la tesis, los cuales cuentan con una traza directa hacia las hipótesis específicas.

1) **Objetivo**

   Evaluar e implementar un software Traductor de Gestos Dinámicos de Brazos que sirva de apoyo ante la carencia de intérpretes frente a la demanda total de su servicio limitado a ciertos horarios.

   Resultado

   Actualmente, a nivel nacional tal como se detalló en capítulos anteriores, existe una carencia de intérpretes frente a la totalidad de personas no oyentes que requieren de sus servicios para interactuar y participar de eventos de forma general. Dado lo anterior y no sólo a nivel local, han aparecido propuestas de índole tecnológico con intenciones de servir de apoyo al rol de intérpretes de lengua de señas. Sin embargo, tal como se revisó en el Estado del Arte, las propuestas tienden a abarcar un alcance definido no siendo soluciones completas,

así, por ejemplo, el desarrollo de un software intérprete de lengua de señas no sólo debe tener en cuenta el desplazamiento de brazos, sino la postura de las manos, expresiones faciales, postura del torso y además debe considerar también la sintaxis, las que pueden variar en menor o mayor grado dependiendo de la zona territorial. Cabe resaltar que un software traductor se limita a no considerar la sintaxis de la lengua de señas a la cual esté enfocada.

En este trabajo, acorde al alcance definido, se abarca el desplazamiento de brazos y manos como paso inicial a la obtención de un intérprete. Según el contexto de aplicación al cual el software se encuentra enfocado, se considera haber satisfecho el primer objetivo puesto que para el conjunto de gestos definidos para su traducción se ha obtenido una elevada tasa de reconocimiento, indicando que el software podrá desempeñarse como apoyo al rol y tareas de los intérpretes, incrementando la cantidad de beneficiarios y permitiendo una mayor interacción y comunicación entre oyentes y no oyentes, lo cual valida la primera hipótesis específica, sin embargo, se recomienda emplearlo una vez finalizado el trabajo futuro especificado.

2) **Objetivo**

Desarrollar e implementar un software económico y usable que facilite a las entidades acatar las normativas de contratación de intérpretes y empleo de medios de interacción y comunicación entre personas oyentes y no oyentes.

Resultado

Primero, se considera necesario resaltar en este punto a no sólo el aspecto económico monetario, sino también al técnico, el cual de la mano con la usabilidad van a permitir a los usuarios contar con menos impedimentos, motivándolos a emplear el software.

Trabajos de terceros usan comúnmente métodos estadísticos (Hidden Markov Models, Redes Neuronales Artificiales, entre otros) como herramientas para entregar soluciones de este tipo, conllevando mayores costos de programación y recursos en general, forzando a los usuarios a realizar repetidas muestras de entrenamiento para permitir al sistema reconocer algún gesto deseado. Dado ello, después de las evaluaciones realizadas en el Estado del Arte, se identificó como alternativa factible emplear DTW al no requerir un entrenamiento exhaustivo, ser de sencilla implementación y producir bajos costos de procesamiento.

Segundo, desde el aspecto monetario, en el Estado del Arte se hizo mención a la existencia de propuestas basadas en guantes de datos y cámaras especializadas de costos elevados en comparación con el sensor postulado en este trabajo, el cual ha tenido gran acogida por diversos investigadores debido a su abaratamiento de precio.

Tal como se detalla en el apartado 5.2.2, para la implementación del software se decidió recurrir a herramientas y tecnologías de acceso gratuito buscando abaratar costos, volviendo viable y factible al proyecto en contraparte a la propuesta o normativa que insta el adiestramiento y contratación de intérpretes como intermediarios y facilitadores. Se puede denotar que dicha normativa es una alternativa con costos (Remuneraciones, pago de viáticos, entre otros) que fácilmente superan nuestra solución de enfoque tecnológico.

Por lo expresado, se considera haber satisfecho el segundo objetivo específico y, además, se llega a validar la segunda hipótesis específica al proponer una solución tecnológica de reducidos costos y pago único, al ser usable y económicamente fácil de adquirir por las entidades/instituciones en su alineamiento con las exigencias de la normativa nacional, promoviendo la accesibilidad, comunicación e interacción de los beneficiarios.

3) **Objetivo**

Enfocar la implementación del software traductor en escenarios de ámbito laboral y nacional, buscando propiciar la Inclusión Social ante la carencia de una educación especializada en sus niveles capaz de insertar a los no oyentes en la sociedad.

Resultado

La necesidad de facilitar medios a las personas no oyentes para relacionarse con la sociedad es un aspecto relevante en la búsqueda de su Inclusión Social. Además, acorde a la normativa internacional y leyes nacionales, las instituciones en general están obligadas a proveer intérpretes o cualquier otra solución que propicie la interacción.

Teniendo en cuenta lo mencionado, se desarrolló la propuesta software enfocada a contextos laborales, de tal forma, una vez concretado el trabajo de etapas futuras, este software propicie la inclusión laboral de quienes requieren comunicarse mediante lengua de señas.

Si bien el software ha sido orientado a ambientes de atención o recepción al público, empleando términos de interacción en esos escenarios, se resalta que su uso es de aspecto general, pudiendo fácilmente adaptarse a cualquier entorno donde se requiera comunicación mediante lengua de señas.

Dado lo anterior, se considera haber satisfecho el tercer objetivo específico. Además, se valida la tercera hipótesis específica dada la posibilidad de emplear el software en diversos contextos de aplicación, en especial el laboral, lo cual permite la contratación de quienes usan lengua de señas para comunicarse, situación la cual propicia la Inclusión Social.

4) **Objetivo**

Implementar y validar el software traductor en tiempo real considerando periodos de procesamiento y tasas de reconocimiento aceptables para facilitar la interacción entre personas oyentes y no oyentes.

Resultado

Tal como se detalló en el Capítulo V, se logró validar la propuesta software tanto en el aspecto de periodos de procesamiento como en las tasas de reconocimiento. Con respecto al primero, se evaluó que el sistema se encuentra apto para entregar respuestas de procesamiento en periodos aceptables hasta con 174 gestos cargados en memoria, después la demora será notable para el usuario. Se estima que la cantidad de 174 gestos permitirá al sistema atender sin mayor inconveniente las traducciones en tiempo real para el contexto de aplicación establecido, por lo cual se considera haber logrado validar y satisfacer el cuarto objetivo específico en este aspecto. Sin embargo, acorde a lo mencionado en ítems anteriores, teniendo en mente la envergadura y trabajo futuro para este proyecto, se cree conveniente realizar una revisión de los métodos empleados con la intención de optimizar y reducir los tiempos.

Para la tasa de reconocimiento, se logró validar la solución obteniendo un 98.18% de eficacia en reconocimiento de gestos. El porcentaje ha sido calculado mediante promedio aritmético considerando cada uno de los gestos definidos. El resultado obtenido supera los mostrados en el Estado del Arte, consecuencia de la configuración o parametrización de los datos obtenidos al realizar la captura de datos durante la ejecución de algún gesto frente al sensor. Por otro lado, los gestos fueron definidos enfocándonos en el contexto de aplicación centrado en la atención al público en una institución educativa.

Acorde a la tasa de reconocimiento obtenida, se considera haber satisfecho el cuarto objetivo específico. Además, se ha logrado validar la cuarta hipótesis específica al implementar un software traductor capaz de responder en tiempos aceptables, facilitando la interacción y comunicación de los beneficiarios.

5.3.3 Validación de la Hipótesis

Citando las encuestas mostradas en los Anexos, a nivel nacional no existen condiciones de equidad en oportunidades para las personas no oyentes [Anexo 02], reflejado en menor acceso a la educación [Anexo 04], reducida participación en la sociedad [Anexo 05] e inserción laboral [Anexo 06], aspectos los cuales ahondan las diferencias de Inclusión Social. Por lo cual, haber logrado la implementación de un software capaz de apoyar a los intérpretes en sus tareas en diversos horarios, y que además, resulte sencillo de emplear, accesible económicamente, apto de ser usado en cualquier contexto laboral generando empleabilidad, y entregando respuestas en tiempo real de forma fluida y eficiente, lo convierte en una propuesta que será fácilmente aceptada y adquirida por las entidades, propiciando una mejor comunicación e interacción en los contextos donde se emplee, generando oportunidades laborales y convirtiéndose en un paso más para el logro de la Inclusión Social.

En congruencia a lo expresado y los resultados mostrados en el punto anterior, el desarrollo de la propuesta, bajo los alcances definidos, ha satisfecho los objetivos e hipótesis específicas, como consecuencia, se ha validado la Hipótesis General de la Tesis.

CAPÍTULO VI: CONCLUSIONES Y TRABAJO FUTURO

El planteamiento del problema, definición de la solución propuesta, la implementación y evaluación han dejado conclusiones que resultan idóneos especificarlos. De igual forma se detalla el trabajo futuro a realizar en la continuación de este proyecto.

6.1 Conclusiones

1) Acorde al área de aplicación y alcance, se ha logrado reconocer el desplazamiento de brazos y manos, obteniendo una elevada tasa de reconocimiento para el conjunto de gestos definidos. Se ha desarrollado un software traductor capaz de apoyar a los intérpretes en su rol, convirtiéndose en un paso inicial para la implementación futura de un software intérprete, sin embargo, se recomienda su empleo una vez finalizadas las siguientes etapas.
2) Frente a propuestas de terceros, en nuestro trabajo se redujeron costos al emplear herramientas de licencia gratuita y económicamente accesibles, convirtiéndose en una propuesta atractiva y económica para el alineamiento de las entidades con las normativas que promueven la Inclusión Social.
3) Nuestra propuesta es capaz de adaptarse a cualquier entorno. Sin embargo, dado el alcance, fue enfocada a contextos laborales de atención al público empleando y reconociendo correctamente términos de uso común en conversaciones de dichos escenarios, de tal forma propiciar la inserción e inclusión laboral de quienes se comunican mediante lengua de señas.
4) La propuesta es capaz de entregar resultados en periodos aceptables hasta con 174 gestos. Dicha cantidad permitirá al sistema atender sin mayor inconveniente y de forma fluida las traducciones en tiempo real. Sin embargo, considerando escalabilidad en etapas futuras, es convenientemente realizar la optimización de los métodos empleados.

6.2 Trabajos Futuros

La presente tesis tiene como objetivo a largo plazo desarrollar un Software Intérprete de Lengua de Señas Peruana. Se han planteado las bases para lograr ello proponiendo la traducción de gestos basándose en la trayectoria realizada por las manos y brazos durante un periodo determinado.

Para lograr el objetivo final se tienen los siguientes puntos como trabajo a desarrollarse en etapas futuras:

1) Un intérprete de lengua de señas requiere el reconocimiento de las posturas de manos y torso, así como las expresiones faciales realizadas durante la ejecución de algún gesto, por lo cual se deben implementar métodos y algoritmos que permitan el seguimiento de las posturas de manos y torso en cada momento durante la realización de los gestos, y también la expresión del rostro.
2) Un software intérprete debe tener en cuenta la traducción de gestos considerando la sintaxis de la lengua de señas empleada para reconocer expresiones e ideas de forma continua acorde al contexto, dejando de traducir palabras independientes.
3) Ampliar el Diccionario de Gestos, el cual estará disponible en internet, así los usuarios a nivel nacional podrán emplear el software sin realizar entrenamiento al sistema.
4) Desarrollar un módulo secundario el cual permitirá traducir palabras escritas o habladas a gestos representados por un avatar, permitiendo a la persona no oyente entender lo que la otra parte está intentando comunicar, por lo cual se logrará una comunicación más fluida.

REFERENCIAS BIBLIOGRÁFICAS

- **REFERENCIAS WEB**

[1] Gutierrez Gamarra Marilia, Lengua de Señas en el Perú, UNIFE, http://es.scribd.com/doc/56770237/Lenguaje-de-Senas-en-el-Peru#, accedido el 25 de Noviembre del 2014.

[2] Estados Partes de la Asamblea General (2006), Convención sobre los Derechos de las Personas con Discapacidad, Naciones Unidas, http://www.un.org/esa/socdev/enable/rights/convtexts.htm, accedido el 20 de Septiembre del 2014.

[3] World Federation of the Deaf (2014), http://wfdeaf.org/, accedido el 20 de Septiembre del 2014.

[5] Pontificia Universidad Católica del Perú (2014), La Lengua de Señas es un Derecho, PUCP, http://blog.pucp.edu.pe/item/145951/la-lengua-de-senas-es-un-derecho, accedido el 20 de Septiembre del 2014.

[6] Dirección General de Educación Básica Especial, Guía para el Aprendizaje de la Lengua de Señas Peruana y Vocabulario Básico, Ministerio de Educación, https://es.scribd.com/doc/56770237/Lenguaje-de-Senas-en-el-Peru, accedido el 25 de Septiembre del 2014.

[7] Instituto Nacional de Estadística e Informática (2013), Primera Encuesta Nacional Especializada sobre Discapacidad, INEI, http://www.conadisperu.gob.pe/encuesta_inei/Resultados%20I%20Encuesta%20Nacional%20de%20Discapacidad%202012.pdf, accedido el 01 de Octubre del 2014.

[8] Asociación de Intérpretes y Guías Intérpretes de Lengua de Señas del Perú, Segundo Encuentro de Intérpretes (2013), ASISEP, http://publimetro.pe/actualidad/noticia-500-mil-sordos-hay-peru-y-solo-18-interpretes-lenguaje-senas-17030?ref=ecr, accedido el 05 de Octubre del 2014.

[9] Asociación de Intérpretes y Guías Intérpretes de Lengua de Señas del Perú (2012), http://www.andina.com.pe/agencia/noticia-menos-del-10-sordos-peru-cuenta-interprete-lenguaje-senas-397215.aspx, accedido el 09 de Octubre del 2014.

[32] XatakaWindows, La Evolución de Kinect y la Importancia Real de Microsoft Research, Xakata, http://www.xatakawindows.com/xbox/la-evolucion-de-kinect-y-la-importancia-de-microsoft-research, accedido el 06 de Diciembre del 2015.

[33] The Imaginative Universal, Quick Reference: Kinect 1 vs. Kinect 2, The Imaginative Universal, http://www.imaginativeuniversal.com/blog/post/2014/03/05/Quick-Reference-Kinect-1-vs-Kinect-2.aspx, accedido el 10 de Diciembre del 2015.

[58] XBOX, La Experiencia XBOX más Rápida y Social de la Historia, https://www.xbox.com/es-ES/xbox-one/experience?xr=shellnav, accedido el 24 de Abril de 2016.

[59] ACM, The 2012 ACM Computing Classification System, https://www.acm.org/publications/class-2012, accedido el 13 de Mayo del 2018.

[60] ACM, The 2012 ACM Computing Classification System – Introducion, https://www.acm.org/publications/class-2012-intro, accedido el 13 de Mayo del 2018.

[61] Asociación de Intérpretes y Guías del Perú (ASISEP), Actitudes de Sordos Adultos y Oyentes con Relación a la Inclusión Social de la Persona Sorda, http://siep.org.pe/archivos/up/10.ppt, accedido el 20 de Octubre del 2014.

- **TESIS**

[4] Ying Yin, Real-time Continuous Gesture Recognition for Natural Multimodal Interaction, Phd Thesis, Massachusetts Institute of Technology, EE.UU, 2014. Paper ISSN: 1943-6092. Paper ISBN: 978-1-4799-4035-6.

[10] Cao Dong, American Sign Language Alphabet Recognition Using Microsoft Kinect, Msc Thesis, Missouri University of Science and Technology, EE.UU, 2015. Paper ISSN: 2160-7516. Paper ISBN: 978-1-4673-6759-2.

[26] Iñaki Iralde Lorente y Alfredo Pina Calafi, Desarrollo de Aplicaciones con Microsoft Kinect, Tesis de Ingeniería, Universidad de Cantabria, España, 2012.

[30] Ahmad Manasrah, Human Motion Tracking for Assisting Balance Training and Control of a Humanoid, Msc Thesis, University of South Florida, EE.UU, 2012.

[31] Daniel Martinez Capilla, Sign Language Translator using Microsoft Kinect XBOX 360™, Tesis de Maestría, The University of Tenessee, EE.UU., 2012.

[40] Rodolgo Javier Galvez Meza, Plataforma Computacional de Captura y Representación del Movimiento en 3D para Apoyo a la Rehabilitación de la Marcha, Tesis de Ingeniería, Pontificia Universidad Católica del Perú, Perú, 2015.

[41] Emre Isikligil, A Method for Isolated Sign Recognition with Kinect, Msc Thesis, Middle East Technical University, Turquía, 2014.

[46] Yi Li, Hand Gesture Recognition Using Kinect, Msc Thesis, University of Louisville, EE.UU, 2012. Paper ISBN: 978-1-4673-2008-5.

[55] Jie Xu, Design and Realization of the Gesture-Interaction System Based on Kinect, Tesis de Ingeniería, UPPSALA Universitet, 2014.

- **PAPERS**

[27] Hee-Deok Yang, Sign Language Recognition with the Kinect Sensor Based on Conditional Random Fields, Sensors 2015, v. 15, pp. 135-147, 2015. ISSN: 1424-8220.

[34] Meinard Müller, Dynamic Time Warping, Information Retrieval for Music and Motion, pp. 318, 2007. ISBN: 978-3-540-74047-6.

[35] Hyo-Rim Choi y Tae-Yong Kim, Directional Dynamic Time Warping for Gesture Recognition, ICMSSP, pp. 22-25, 2017. ISBN: 978-1-4503-5314-4.

[36] Peng Wang, Haixun Wang y Wei Wang, Finding Semantics in Time Series. SIGMOD, pp. 385-396, 2011. ISBN: 978-1-4503-0661-4.

[37] Jing Jiang y Chengxiang Zhai, Extraction of Coherent Relevant Passages Using Hidden Markov Models, ACM Transactions on Information Systems, v. 24, pp. 295-319, 2006.

[38] Farhad Bulbul, Yunsheng Jiang y Ma Jinwen, DMMs-Based Multiple Features Fusion for Human Action Recognition, International Journal of Multimedia Data Engineering and Management (IJMDEM), v. 6, n. 4, pp. 23-39, 2015. ISSN: 1947-8534.

[39] Feng Jiang, Shengping Zhang, Shen Wu, Yang Gao y Debin Zhao, Multi-layered Gesture Recognition with Kinect, Journal of Machine Learning Research, v. 16, pp. 227-254, 2015. ISSN: 1532-4435.

[42] Paul Doliotis, Alexandra Stefan, Christopher McMurrough, David Eckhard y Vassilis Athitsos, Comparing Gesture Recognition Accuracy Using Color and Depth Information, 2011. ISBN: 978-1-4503-0772-7.

[43] André Alexandros Chaaroui, Pau Climent-Pérez y Francisco Florez-Revuelta, A Review on Vision Techniques Applied to Human Behavior Analysis for Ambient-Assisted Living, International Journal of Expert Systems with Applications, v. 39, n. 12, pp. 10873-10888, 2012. ISSN: 0957-4174.

[44] Heng Wang y Cordelia Schmid, Action Recognition with Improved Trajectories, In Proceedings of the IEEE International Conference on Computer Vision, v. 1, n. 1, pp. 3551-3558, 2013. ISBN: 978-1-4799-2840-8.

[45] Lulu Chen, Hong Wei y James Ferryman, A Survey of Human Motion Analysis Using Depth Imagery, Pattern Recognition Letters, v. 1, n. 1, pp. 1995-2006, 2013. ISSN: 0167-8655.

[47] Chan-Su Lee, Zeungnam Bien, Gyu-Tae Park, Wong Jang y Jong-Sung Kim, Real-Time Recognition System of Korean Sign Language Based on Elementary Components, In Proceedings of the Sixth International Conference on Fuzzy Systems, v. 3, pp. 1463-1468, 1997. ISBN: 0-7803-3796-4.

[48] Rodrigo Ibañez y Damián Fanaro, Herramienta para Facilitar el Desarrollo de Aplicaciones Basadas en Kinect, EST, v. 16, pp. 321, 2013. ISSN: 1850-2946.

[49] Kenji Oka, Yoichi Sato y Hideki Koike, Real-Time Fingertip Tracking and Gesture Recognition, IEEE Computer Graphics and Applications, v. 22, pp. 64-71, 2002. ISSN: 0272-1716.

[50] Xin Zhao, Xue Li, Chaoyi Pang, Xiaofeng Zhu y Quan Z. Sheng, Online Human Gesture Recognition from Motion Data Streams, Proceedings of the 21st ACM International Conference on Multimedia, pp. 23-32, 2013. ISBN: 978-1-4503-2404-5.

[51] Nikolaos Gkigkelos y Christos Goumopoulos, Greek Sign Language Vocabulary Recognition Using Kinect, Proceedings of the 21st Pan-Hellenic Conference on Informatics, 2017. ISBN: 978-1-4503-5355-7.

[52] Sohaib Laraba, Joelle Tilmanne y Dutoit Thierry, Adaptation Procedure for HMM-Based Sensor-Dependent Gesture Recognition, Conference Paper, 2015. ISBN: 978-1-4503-3991-9.

[53] Swapnil Athavale y Mona Deshmukh, Dynamic Hand Gesture Recognition for Human Computer Interaction: A Comparative Study, International Journal of Engineering Research and General Science, v. 2, n. 2, 2014. ISBN: 2091-2730.

[54] Josep Maria Carmona y Joan Climent, A Performance Evaluation of HMM and DTW for Gesture Recognition, CIARP 2012, v. 1, pp. 236-243, 2012. ISBN: 978-3-642-33274-6, ISSN: 0302-9743.

[56] Verónica Lopez-Ludeña, Rubén San-Segundo, Carlos González, Juan Carlos López, y José Pardo, Methodology for Developing a Speech Sign Language Translation System in a New Semantic Domain, ADFA, v. 1, pp. 1, 2011.

- **LIBROS**

[11] McNeill David, Language and Gesture, University of Cambridge, 2000. ISBN 0-521-77166-8.

[12] England David, Whole Body Interaction, Springer, 2011. ISBN 978-0-85729-432-6. ISSN 1571-5035.

[13] McNeill David, Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press., 1992. ISBN 0-226-56132-1.

[14] Ghaoui Claude, Encyclopedia of Human Computer Interaction, Liverpool John Moores University, 2006. ISBN 1-59140-798-2.

[15] Jounghyun Kim Gerard, Human-Computer Interaction: Fundamentals and Practice, CRC Press, 2015. ISBN 978-1-4822-3390-2

[16] Juran J. M., Juran y la calidad por el diseño, Juran Institute, 1992. ISBN: 84-7978-215-3.

[17] Pallás Areny Ramón, Sensores y Acondicionadores de Señal, Universitat Politecnica de Catalunya, 2003, ISBN: 84-267-1344-0.

[18] Vetelino John, Reghu Aravind, Introduction to Sensors, CRC Press, 2010. ISBN 978-1-4398-0852-8

[19] Wilson Jon S., Sensor Technology Handbook, Newnes, 2005. ISBN 0-7506-7729-5

[20] Li Stan Z., Jain Anil K., Encyclopedia of Biometrics, 2009. ISBN: 978-0-387-73002-8.

[21] Henry Peter, Experimental Robotics, The 12th International Symposium on Experimental Robotics ISER, 2010. ISBN: 978-3-642-28572-1.

[22] Thierry Bouwmans, Porikli Fatih, Benjamin Höferlin, Antoine Vacavant, Background Modeling and Foreground Detection for Video Surveillance, CRC Press, 2015. ISBN: 978-1-4822-0538-1.

[23] Shao Ling, Han Jungong, Kohli Pushmeet, Zhang Zhengyou, Computer Vision and Machine Learning with RGB-D Sensors, Springer, 2014. ISSN: 2191-6586, ISBN: 978-3-319-08650-7.

[24] Salazar Sutil Nicolás, Motion and Representation: The Language of Human Movement, Massachusetts Institute of Technology, 2015. ISBN: 978-0-262-02888-2.

[25] Sarbolandi Hamed, Lefloch Damien, y Kolb Andreas, Computer Vision and Image Understanding, ELSEVIER, 2015. ISSN: 1077-3142.

[28] Borenstein Greg, Making Things See, Maker Media, 2012. ISBN: 978-1-449-30707-3.

[29] Kramer Jeff, Burrus Nicolas, Echtler Florian, Herrera Daniel, Parker Matt, Hacking the Kinect, Apress, 2012. ISBN: 978-1-4302-3868-3.

[57] Dirección General de Educación Básica Especial, Lengua de Señas Peruana, Ministerio de Educación, v. 1, pp. 1, 2010.

ANEXOS

Anexo 01:

"Primera Encuesta Nacional Especializada sobre Discapacidad"

INEI – 2012

Individuos en total: 532 209 personas

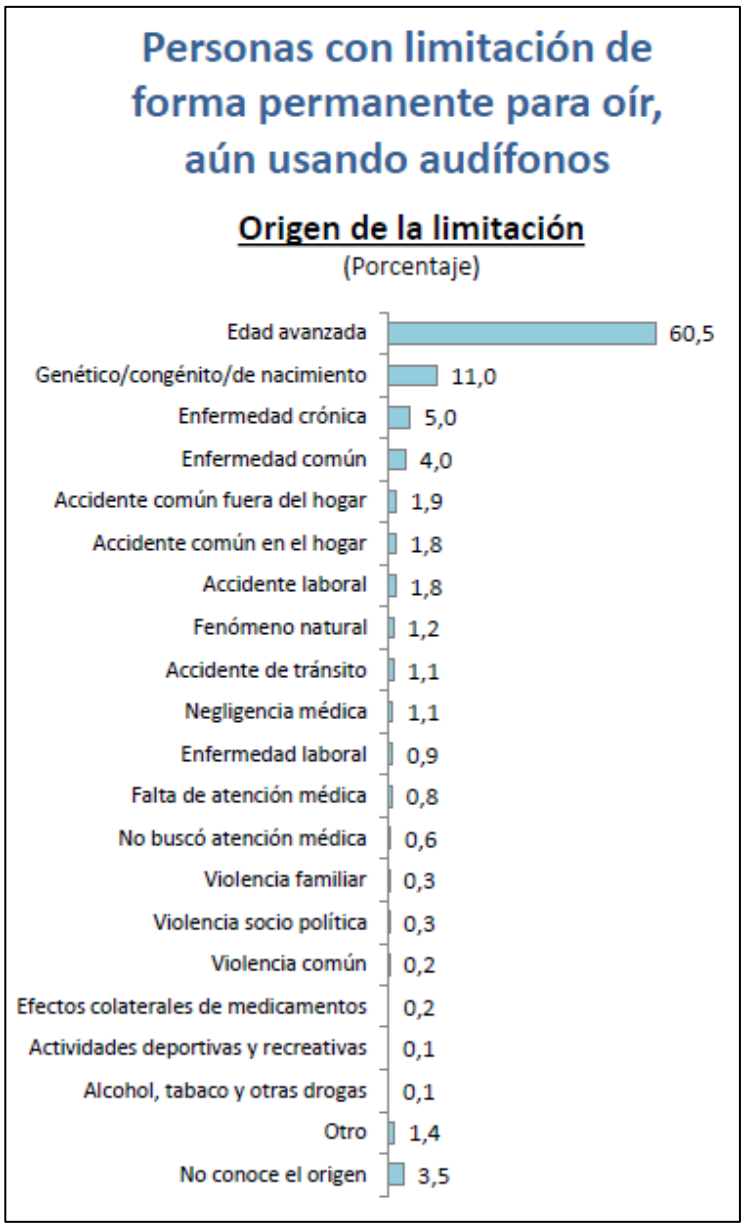


Figura 6.1 Personas con Limitación de Forma Permanente para Oír, Origen de la Limitación. Fuente: [7]

Anexo 02:

"Opiniones de sordos y oyentes con relación a la inclusión social de la persona sorda"

I Seminario Nacional de Investigación Educativa - SIEP - 2006

¿En el país existen condiciones de equidad, igualdad de oportunidades y trato digno a las personas sordas adultas que trabajan?

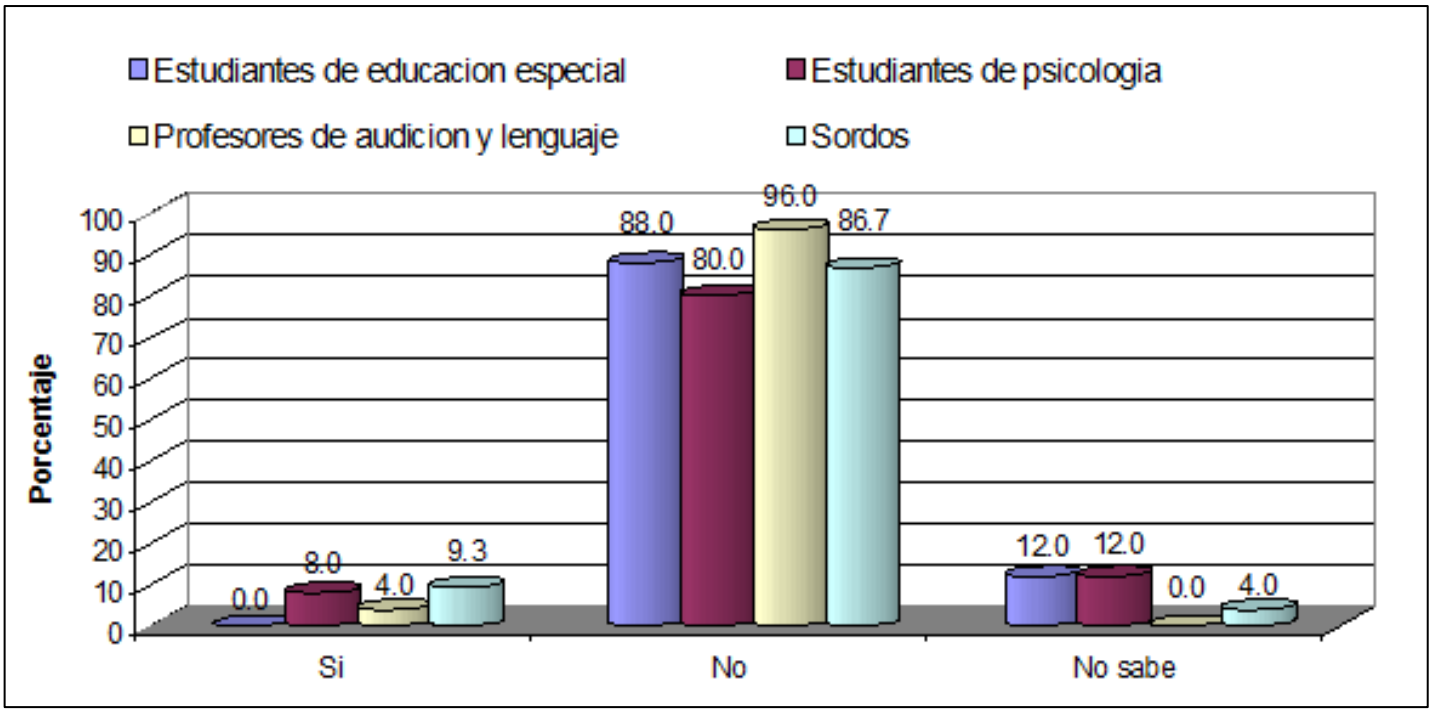


Figura 6.2 Opiniones de Sordos y Oyentes con Relación a la Inclusión Social de la Persona Sorda.

Fuente: [61]

Anexo 03:

"Primera Encuesta Nacional Especializada sobre Discapacidad"

INEI – 2012

Individuos en total: 532 209 personas

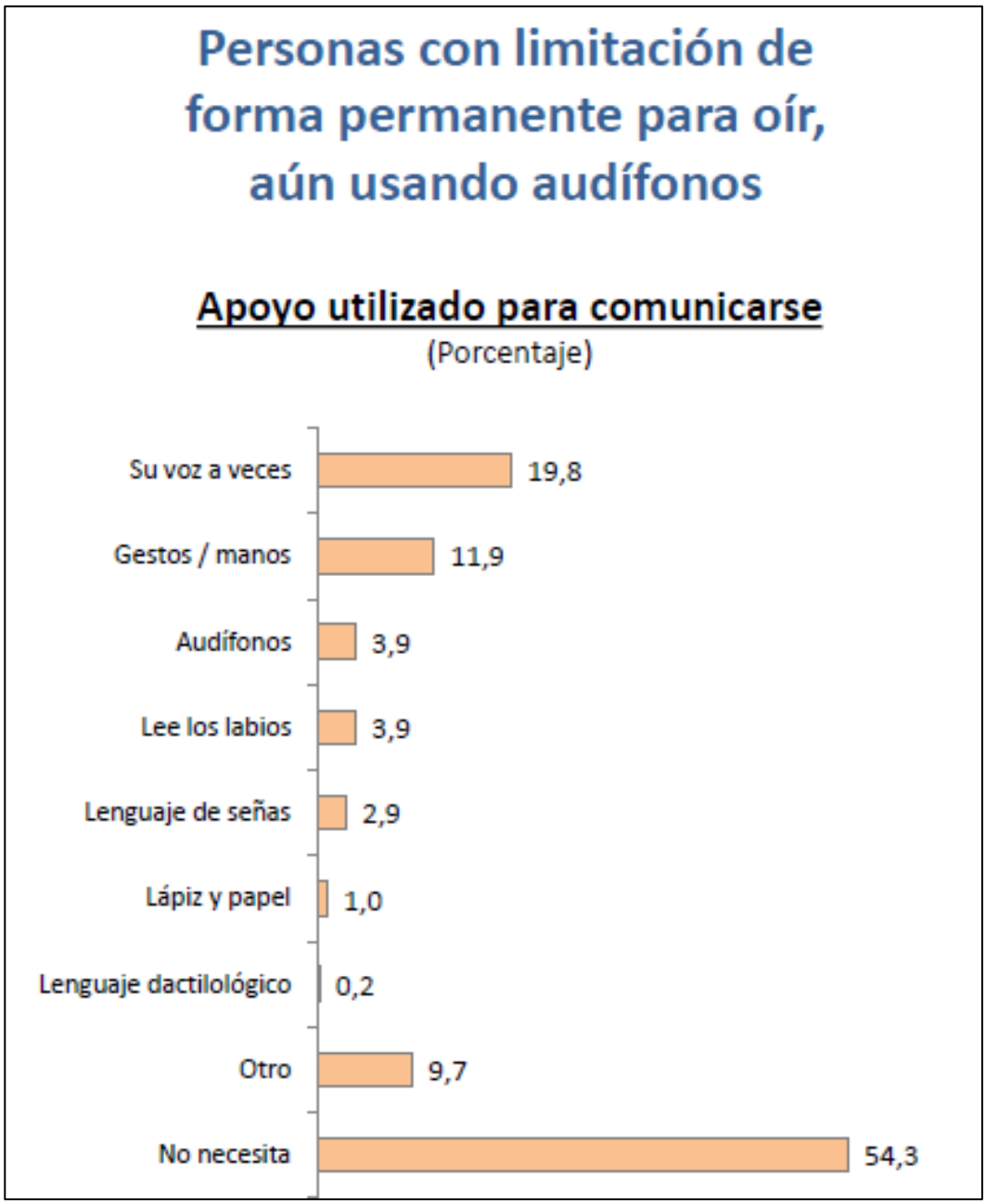


Figura 6.3 Personas con Limitación de Forma Permanente para Oír, Apoyo utilizado para Comunicarse.

Fuente: [7]

Anexo 04:

"Primera Encuesta Nacional Especializada sobre Discapacidad"

INEI – 2012

Individuos en total: 1 575 402 personas

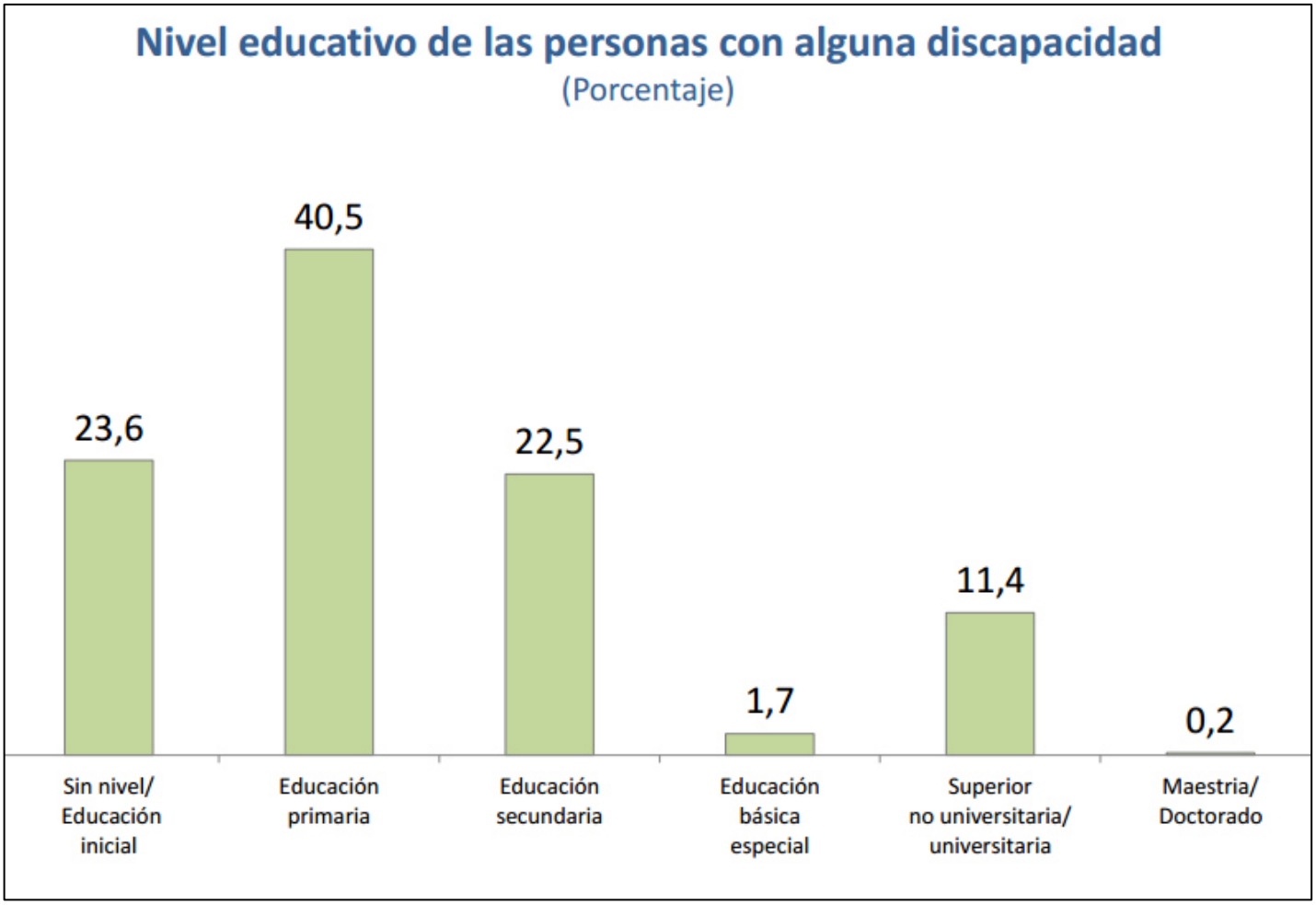


Figura 6.4 Nivel Educativo de las Personas con Alguna Discapacidad. Fuente: [7]

Anexo 05:

"Opiniones de sordos y oyentes con relación a la inclusión social de la persona sorda"

I Seminario Nacional de Investigación Educativa - SIEP - 2006

¿Se toma en cuenta la opinión de las personas sordas adultas para garantizar el ejercicio de sus derechos civiles?

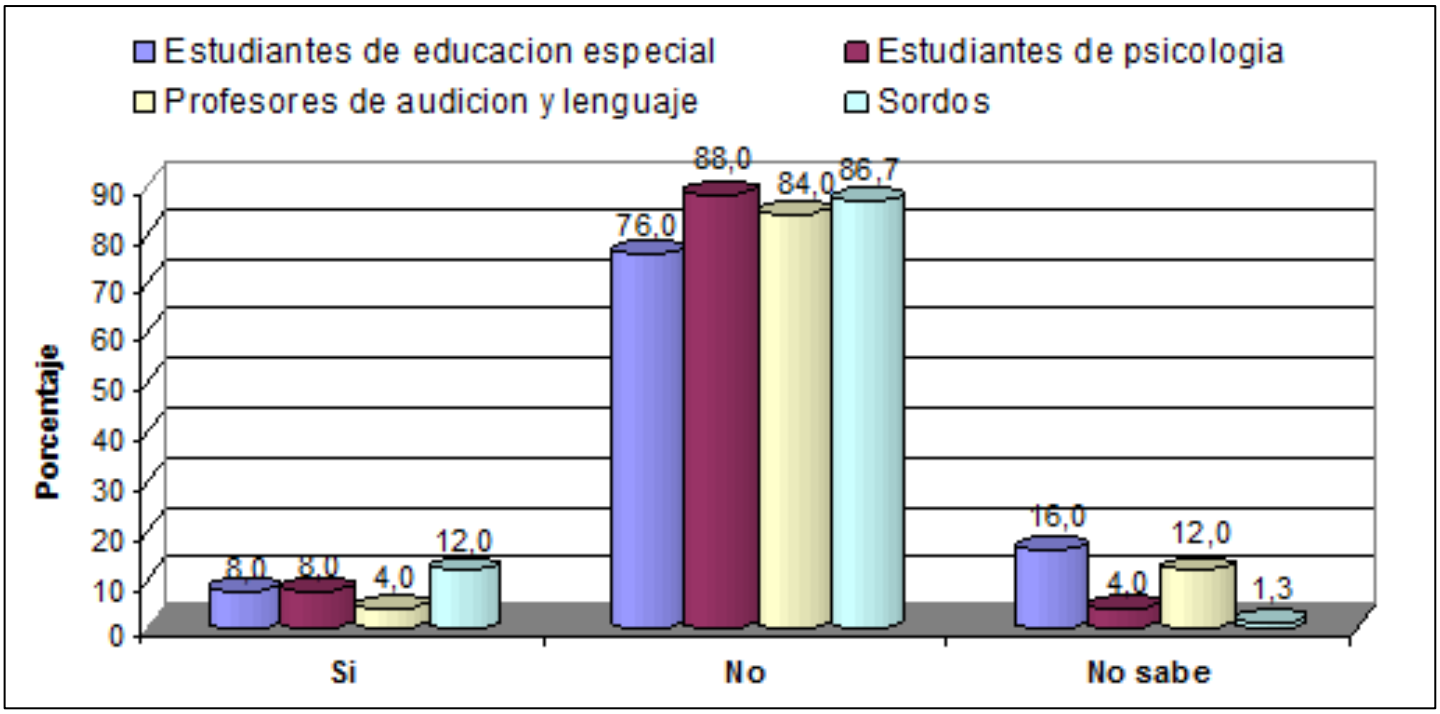


Figura 6.5 Opiniones de Sordos y Oyentes con Relación a la Inclusión Social de la Persona Sorda.

Fuente: [61]

Anexo 06:

"Primera Encuesta Nacional Especializada sobre Discapacidad"

INEI – 2012

Individuos en total: 1 456 543 personas

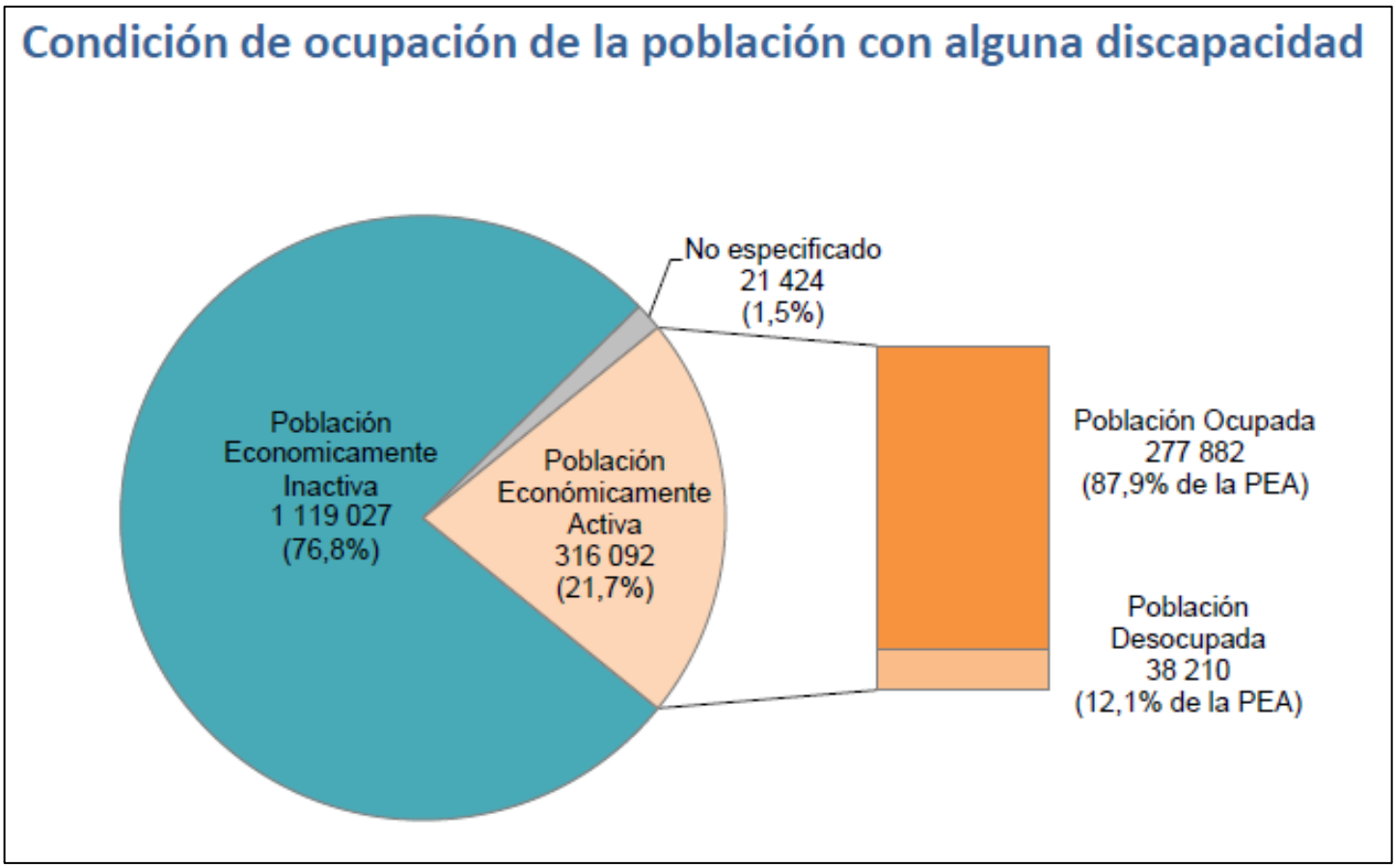


Figura 6.6 Condición de Ocupación de la Población con Alguna Discapacidad. Fuente: [7]

Anexo 07: Requerimientos Técnicos

Para la implementación del software, se considera recomendable establecer ciertos requerimientos técnicos mínimos para su correcto funcionamiento:

- Requerimientos Hardware:
  - Sensor Kinect XBOX 360™.
  - Adaptador USB del Sensor Kinect XBOX.
  - Puerto USB 2.0 dedicado.
  - Procesador Core i3.
  - Memoria RAM de 4GB.
- Requerimientos Software:
  - Windows 7 o superior.
  - SDK 1.7 del sensor Kinect.
  - .NET Framework 3.5

Anexo 08: Interfaces

- Interface Principal

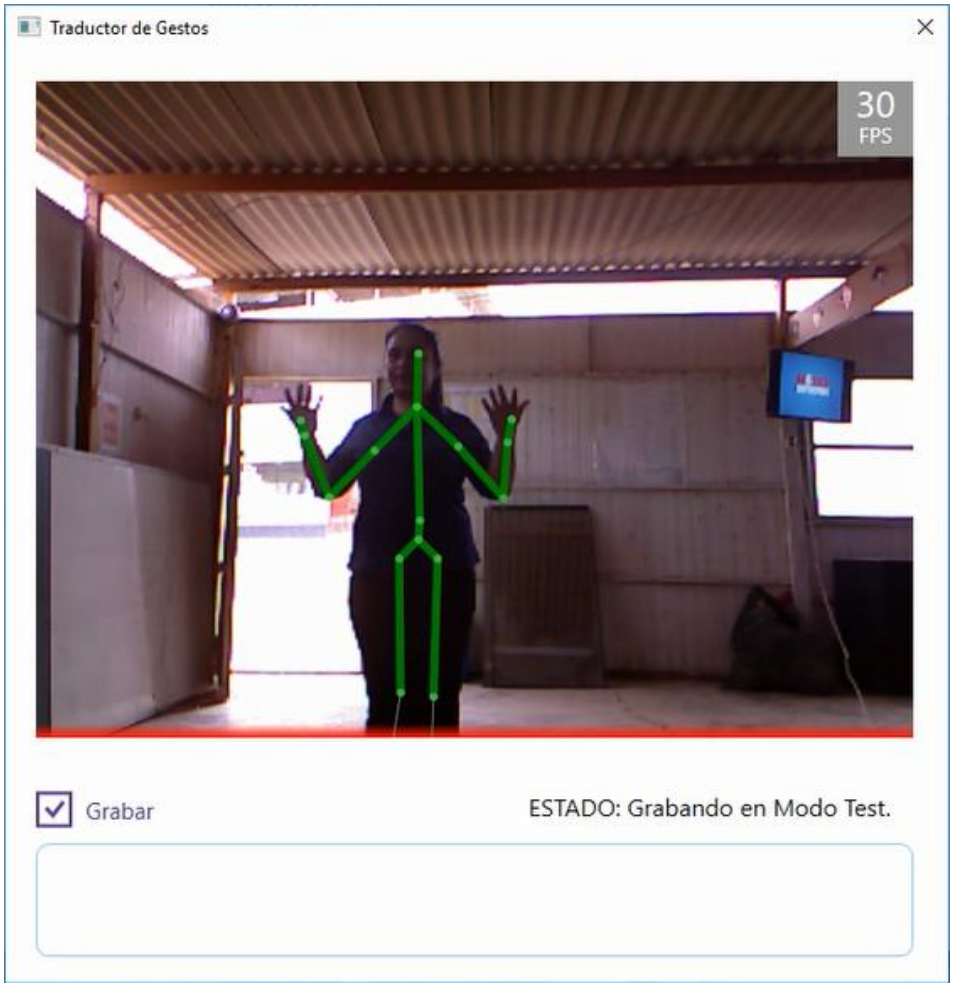


Figura 6.7 Interfaz Principal del Software Propuesto. Fuente: Elaboración propia.

- Interface de Configuración

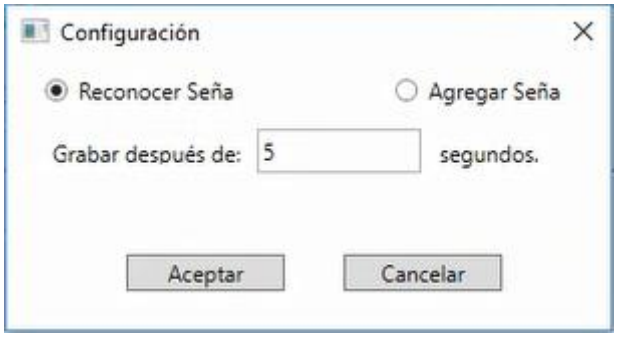


Figura 6.8 Interfaz de Configuración del Software Propuesto. Fuente: Elaboración propia.

Anexo 09: Manual de Usuario

A continuación, se presenta un ligero Manual de Usuario con la intención de facilitar el uso del sistema. Está dividido en partes acorde a la tarea que se desee realizar.

1) Iniciar la Aplicación
   a) Verifique que el sensor Kinect se encuentre conectado a una fuente de poder eléctrico y también a la computadora mediante el conector USB. Si todo es correcto, usted podrá visualizar un LED amarillo parpadear periódicamente en el sensor.
   b) Una vez instalada la aplicación en la computadora, identifique el ícono de la aplicación y haga doble clic sobre la misma. Debe esperar unos segundos.
   c) En caso el software haya cargado correctamente, le aparecerá una interface como la siguiente:

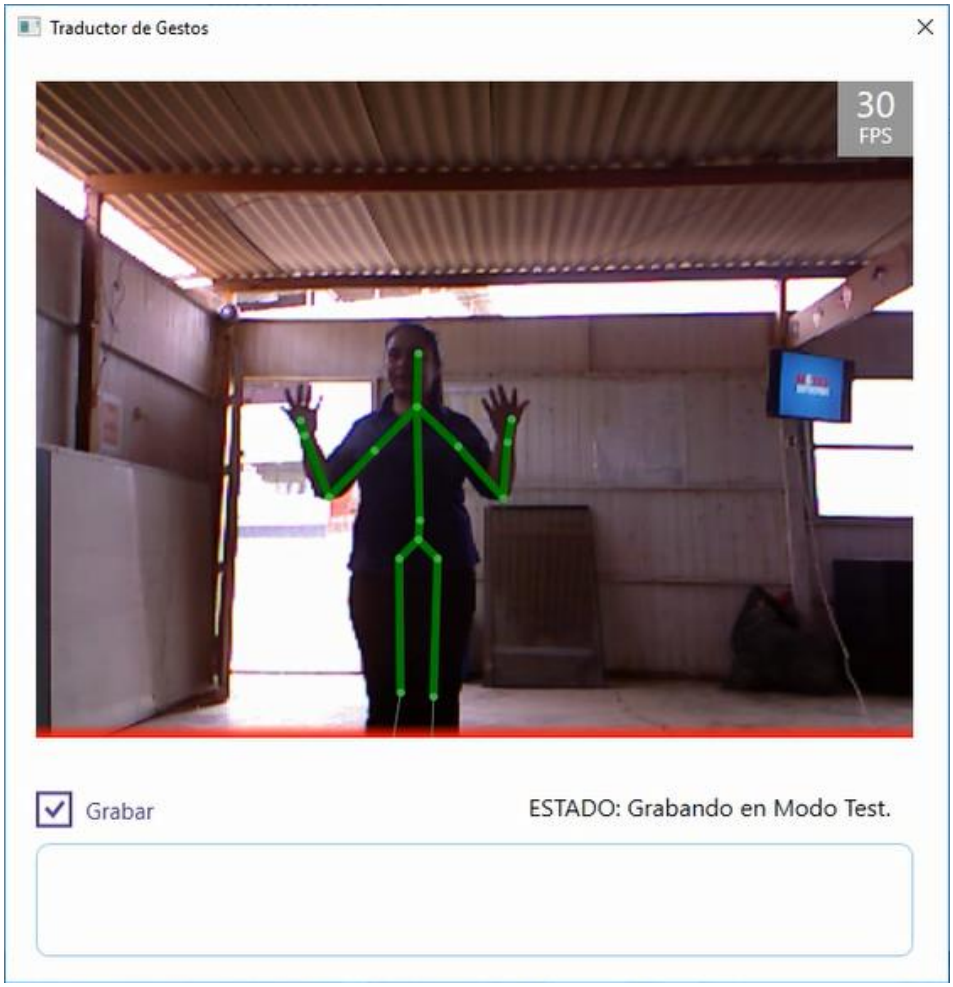


Figura 6.9 Interface inicial de la propuesta software. Fuente: Elaboración propia.

En la imagen se observa una persona realizando un gesto cualquiera, por lo cual, como una última validación se recomienda a usted pararse frente al sensor y esperar que el sistema dibuje su silueta con trazos de color verde la cual se desplazará con usted en tiempo real.

En la última imagen se pueden identificar secciones dentro, las cuales serán detalladas:

a) Área de Visualización: Dada por el bloque más grande, lugar donde se muestra en video lo que el sensor Kinect está capturando en ese instante mediante sus cámaras. Si usted se posiciona frente podrá observar una silueta constituida por trazos de color verde que se mueve acorde a usted.
b) Grabar: Representado por un cuadro color morado donde usted puede hacer clic. Tiene por finalidad abrir una pequeña ventana con la cual podrá indicar al sistema si desea almacenar un nuevo gesto para su posterior reconocimiento, o si sólo desea que el sistema reconozca el movimiento que usted va a realizar frente al sensor.
c) Área de Estado: Un pequeño texto mostrado al lado derecho del cuadro Grabar. Tiene función informativa y sus valores pueden variar acorde a las acciones del usuario, así por ejemplo se tiene:
   - "Espera…", cuando el sistema se encuentra a la espera de las indicaciones del usuario.
   - "Grabando en Modo Test", cuando el sistema se encuentra almacenando los datos provenientes del gesto que se encuentra realizando el usuario frente al sensor.
   - "Se comenzará a grabar en X segundos", indica la cantidad de segundos pendientes antes de iniciar la grabación de los datos provenientes del sensor.
d) Área de Resultados: Ubicado en la parte inferior, es el último bloque de la interfaz mostrada. Muestra la información correspondiente al resultado del procesamiento una vez que el usuario ha realizado algún gesto frente al sensor en Modo Reconocimiento.

2) Almacenar un Gesto

Esta opción se activa haciendo clic en el cuadro Grabar de la imagen anterior.

a) Acto seguido, aparecerá la siguiente ventana con la opción "Agregar Seña" seleccionada y solicitándole ingrese el "Nombre de la Seña" a almacenar.

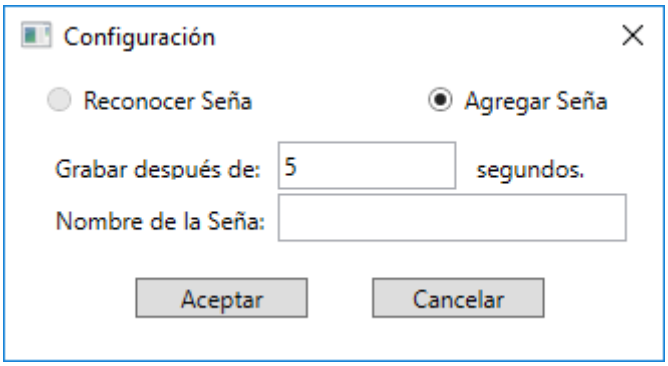


Figura 6.10 Ventana de configuración del software para agregar una seña. Fuente: Elaboración propia.

b) Una vez ingresados los valores, indicar "Aceptar". La ventana de configuración desaparecerá.

c) Acto seguido en la ventana principal aparecerá un contador en cuenta regresiva el cual informará al usuario los segundos restantes para el inicio de la grabación.

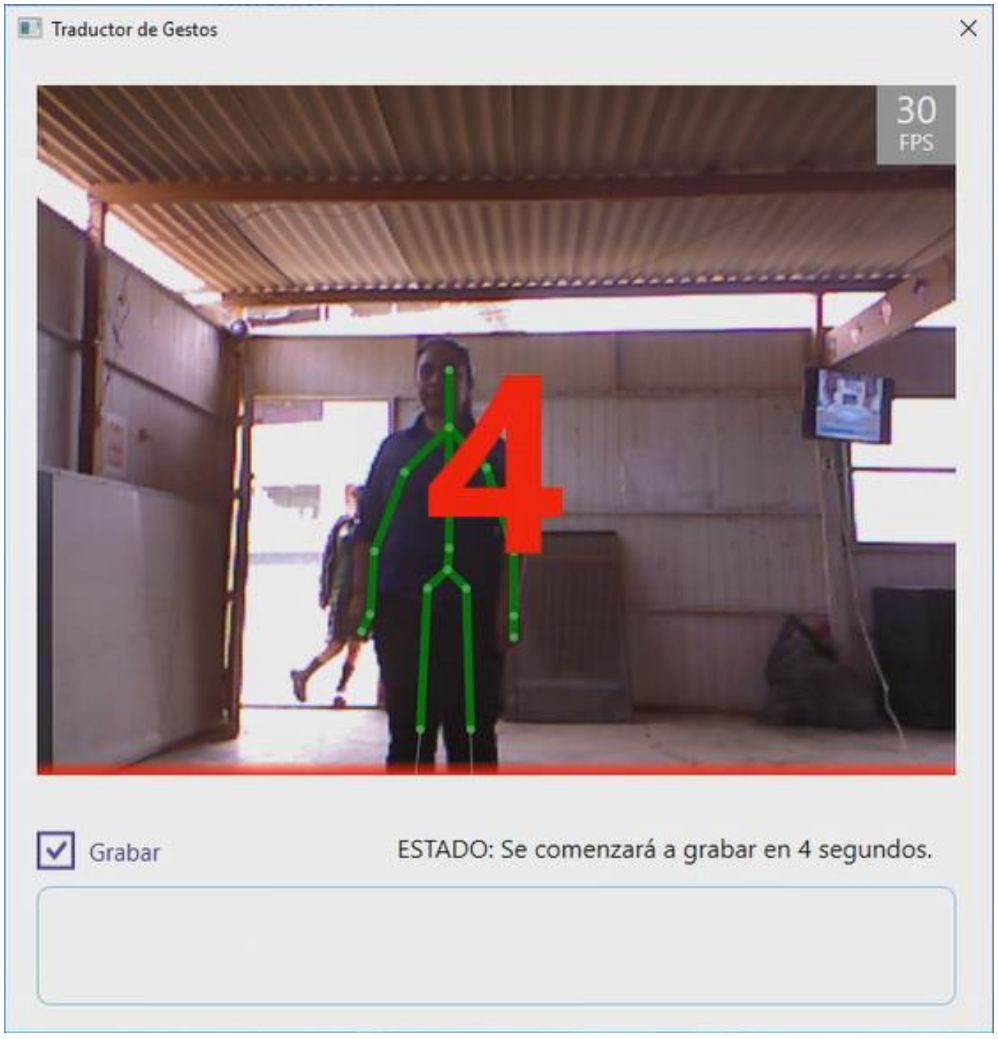


Figura 6.11 Contador regresivo en la interface principal del software. Fuente: Elaboración propia.

3) Reconocer un Gesto

Sólo estará disponible cuando existan gestos almacenados en el diccionario de gestos de la computadora.

Los pasos a realizar son los mismos que el procedimiento indicado en (2) para Almacenar un Gesto, con la diferencia de que en la Ventana de Configuración es necesario seleccionar “Reconocer Seña” e indicar además la cantidad de segundos para iniciar la grabación del gesto.

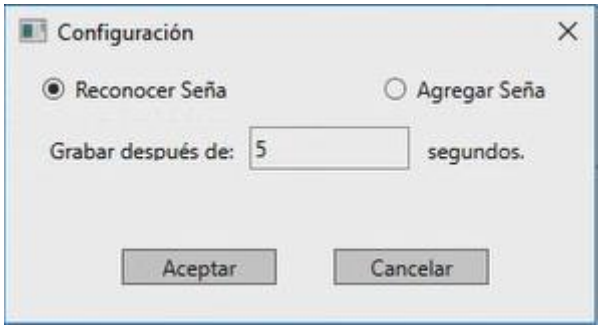


Figura 6.12 Ventana de configuración del software para reconocer una seña. Fuente: Elaboración propia.

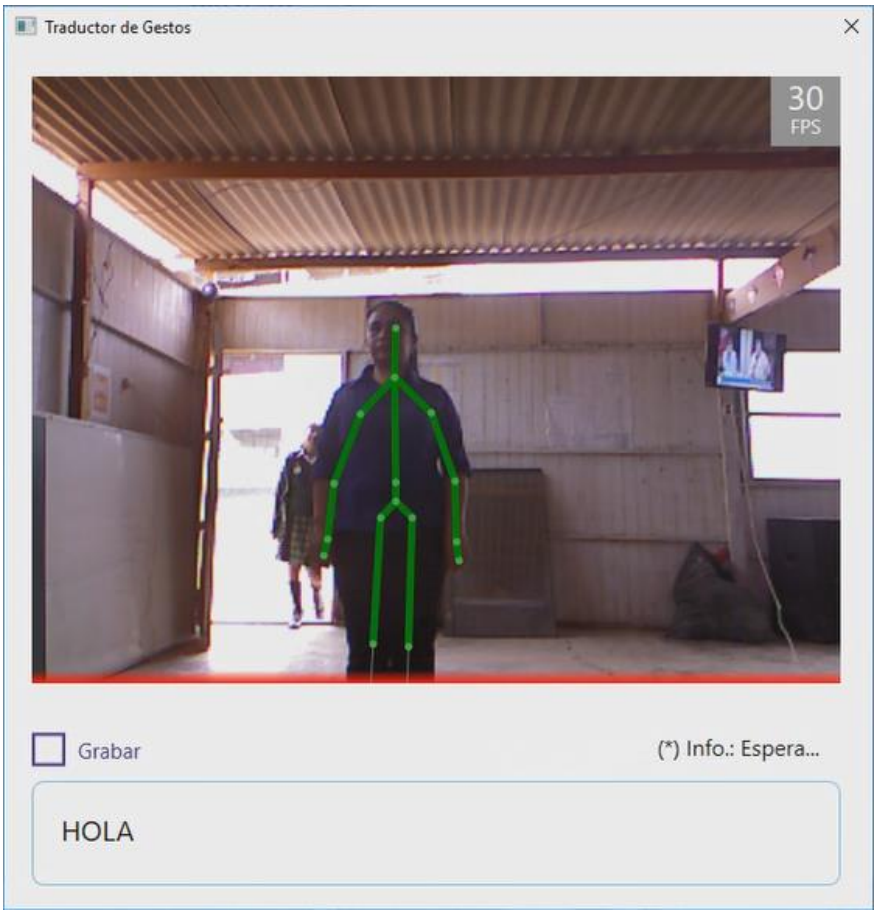


Figura 6.13 Interface principal del software indicando el resultado de procesamiento “Hola”. Fuente: Elaboración propia.

Finalizada la grabación, el sistema pintará la respuesta del procesamiento en el Área de Resultados del formulario principal. En la imagen anterior se observa que el resultado fue "Hola".

Cabe resaltar que tanto para Almacenar como Reconocer un Gesto, el sistema espera que el inicio y finalización de su ejecución considere posicionar sus manos en modo reposo por debajo de la altura de la cintura, así como se observa en la figura anterior.

NOTAS

[Nota01] Entre los más resaltantes, cito:

(c) Reafirmar la universalidad, indivisibilidad, interdependencia e interrelación de todos los derechos humanos y libertades fundamentales, así como la necesidad de garantizar que las personas con discapacidad ejerzan plenamente y sin discriminación.

(h) Reconocer también que la discriminación contra cualquier persona por razón de su discapacidad constituye una vulneración de la dignidad y el valor inherentes del ser humano.

(k) Observando con preocupación que, pese a estos diversos instrumentos y actividades, las personas con discapacidad siguen encontrando barreras para participar en igualdad de condiciones con las demás en la vida social y que se siguen vulnerando sus derechos humanos en todas partes del mundo.

(m) Reconociendo el valor de las contribuciones que realizan y pueden realizar las personas con discapacidad al bienestar general y a la diversidad de sus comunidades, y que la promoción del pleno goce de los derechos humanos y las libertades fundamentales por las personas con discapacidad y de su plena participación tendrán como resultado un mayor sentido de pertenencia de estas personas y avances significativos en el desarrollo económico, social y humano de la sociedad y en la erradicación de la pobreza.

[Nota02] Un intérprete necesita tener conocimientos y una preparación superior para poder desarrollar su labor. Ello deriva del oficio de estas personas que consiste en propiamente ser "Intérprete" de dos lenguajes, la lengua hablada, perteneciente a la comunidad oyente, el cual posee una gramática y sintaxis propia, y la lengua de señas que también posee una gramática y sintaxis propia. Dicha situación se agrava más si se hace hincapié en las variaciones que puedan tener ambos según el lugar geográfico en el que se utilicen.

[Nota03] Entre los beneficios se puede destacar:

- ✓ Registro oficial de intérpretes.
- ✓ Reconocimiento oficial de la carrera como profesión, lo que no se da por la falta de institucionalización.
- ✓ Aumento de la comunidad de intérpretes y mayor cobertura de la población con sordera, lo que garantizaría mayor Inclusión Social.
- ✓ Capacitación oficial de calidad y certificación.

ÍNDICE ANALÍTICO