
Segmentacion de manos basada en Redes Neuronales.

Su uso en un sistema embebido para control de TV.

Por

CARLA ELIZABETH LUNA GENNARI



Facultad de Informática
UNIVERSIDAD NACIONAL DE LA PLATA

Directora: Dra. Laura Lanzarini
Co-Director: Lic. César Estrebou

TESINA DE LICENCIATURA EN SISTEMAS
DICIEMBRE 2017

TABLA DE CONTENIDOS

	Pagina
Lista de Figuras	5
1 Reconocimiento de Gestos	11
1.1 Introducción	11
1.2 Qué son los gestos	12
1.3 Clasificación de los Gestos	13
1.3.1 Clasificación de Efrón	13
1.3.2 Clasificación de Kendon	15
1.3.3 Clasificación de Mc Neill	15
1.4 Clasificaciones más recientes	16
1.4.1 Wexelblat	16
1.4.2 Karam	17
1.5 Taxonomía de los gestos	18
1.6 Clasificación de gestos en HCI	20
1.6.1 Clasificación para su reconocimiento	23
1.6.2 Interfaces Touchless	24
1.6.3 Dispositivos de entrada	25
1.6.4 Algoritmos	26
1.7 Conclusión	29
2 Visión computacional	31
2.1 Introducción	31
2.2 Conceptos básicos	32
2.3 Imágenes digitales	32
2.3.1 RGB	33
2.3.2 Matiz, Saturación, Intensidad	33
2.3.3 Matiz, Saturación, Luminosidad, TSL (Tint, Saturation, Lightness)	35

TABLA DE CONTENIDOS

2.3.4	YCrCb	36
2.3.5	CIE-Lab	37
2.3.6	Comparación de los sistemas de color	37
2.4	Procesamiento de Imágenes	39
2.4.1	Filtros de dominio de frecuencia	39
2.4.2	Filtros Espaciales	40
2.4.3	Morfología de la Imagen	41
2.4.4	Ruidos	43
2.5	Segmentación de objetos	45
2.5.1	Descriptores	47
2.6	Conclusión	48
3	Segmentación de Manos	49
3.1	Introducción	49
3.2	Comunicación hombre-maquina	50
3.3	Productos utilizados para reconocer gestos	51
3.3.1	Kinect- Microsoft	52
3.3.2	RealSense- Intel	53
3.3.3	Leap-Motion	55
3.4	Experiencias en segmentación de manos	56
3.4.1	Marcadores	57
3.4.2	Clasificación para lenguaje de señas	58
3.4.3	Reconocimiento del alfabeto	58
3.5	Conclusión	60
4	Redes Neuronales	61
4.1	Introducción	61
4.2	Definición	61
4.2.1	Redes neuronales artificiales	63
4.2.2	Perceptrón	64
4.2.3	Redes Feedforward	65
4.3	Red neuronal Energía Coulombica Restringida (RCE)	66
4.3.1	Entrenamiento de la red	69
4.3.2	Clasificación	71
4.4	Limitaciones	72
4.5	Conclusión	73

TABLA DE CONTENIDOS

5 Aplicación en control de TV	75
5.1 Introducción	75
5.2 Parte Uno - Hardware	76
5.2.1 Análisis de posibles soluciones	76
5.2.2 Integración	77
5.2.3 Construcción del prototipo	78
5.3 Parte dos - Software	80
5.4 Software para Reconocimiento de Gestos	81
5.4.1 Adquisición	82
5.4.2 Segmentación	83
5.4.3 Extracción de características	84
5.4.4 Detección del gesto	84
5.4.5 Ejecución de Acción	85
5.5 Pruebas/Resultados	86
5.5.1 Sistemas de Color y Red Neuronal RCE	87
5.5.2 Reconocimiento de Gestos	88
5.6 Conclusión	90
6 Conclusiones	93
6.1 Líneas de trabajo futuras	94
A Anexo A - Protocolo de Control Remoto Infrarrojo	97
Bibliografía	101

LISTA DE FIGURAS

FIGURAS	Pagina
1.1 a) Mano cerrada; b) Mano abierta	21
1.2 Reconocimiento de Gestos	24
1.3 Dispositivos de entrada. (a) Guantes con cable. (b) Cámaras de profundidad. (c) Cámaras estéreo. (d) Controladores basados en gestos.	26
2.1 RGB	34
2.2 Modelo de color HSI	35
2.3 Modelo de color TSL	36
2.4 Modelo de color YCbCr	37
2.5 Modelo de color Cie-Lab	38
2.6 Ejemplo Ruido gaussiano	44
2.7 Ruidos - Sal y pimienta	44
2.8 Máscaras de Laplaciano	46
3.1 Kinect de Microsoft	52
3.2 Sensor de profundidad de Kinect	53
3.3 Sistema de imágenes 3D - Real Sense - Intel	54
3.4 Flujo de captura de imagen en profundidad - Real Sense	55
3.5 Leap Motion	56
3.6 Imágenes no segmentadas de la base de datos LSA	59
3.7 Calibración A	60
3.8 Calibración B	60
4.1 Neurona	62
4.2 Arquitectura Perceptrón	64
4.3 Arquitectura Feedforward	65
4.4	67

LISTA DE FIGURAS

4.5	Arquitectura RCE	68
4.6	Muestra de resultados de la ejecución del algoritmo en 3D en el sistema de color RGB	71
5.1	Módulos que componen el hardware. (a) Raspberry PI 3. (b) Cámara web. (c) Sensor de luminosidad. (d) Emisor y receptor infrarrojos.	79
5.2	Configuraciones	80
5.3	Proceso de reconocimiento del gesto y ejecución de comandos.	82
5.4	Imagen procesada con RCE	83
5.5	Mano Recortada	84
5.6	Ejemplo mano abierta	85
5.7	Ejemplo mano cerrada	85
5.8	(a) Subir el volumen. (b) Bajar el volumen. (c) Siguiente canal	86
5.9	Red neuronal RCE en diferentes sistemas de representación del color.	88
5.10	Formas de manos. (a) Mano abierta con dedos separados, (b) Mano abierta con dedos juntos. (c) Mano cerrada con índice y pulgar separados.	88
5.11	Resultados en porcentaje de la detección de los gestos.	89
5.12	Resultados en porcentaje de la detección de los gestos. Prueba 2	90
5.13	Gestos detectados por el algoritmo. (a) Abajo mano abierta. (b) Arriba mano cerrada. (c) Derecha mano abierta. (d) Izquierda mano cerrada.	91
A.1	Cada pulso se enciende y apaga a una frecuencia de 38 kHz	98
A.2	99
A.3	Protocolo IR	100

RESUMEN

Una interfaz hombre-máquina juega un papel importante a la hora de transmitir una intención de un usuario a un dispositivo. Hoy en día, la solución basada en visión ha estado atrayendo mucha atención, ya que no requiere sensores de fijación en el cuerpo.

En la actualidad, el reconocimiento de gestos hechos con las manos se ha convertido en un tema de sumo interés por parte de investigadores que trabajan en áreas muy diferentes como por ejemplo realidad aumentada, control de dispositivos o reconocimiento del lenguaje de señas Argentino entre otras [24, 27, 30]. Se trata de un problema formado por dos partes perfectamente diferenciadas. La primera de ellas es la encargada de segmentar las manos y la segunda se ocupa de la caracterización y reconocimiento del gesto. Ambas presentan diferentes niveles de complejidad según el tipo de cámara que se utilice así como la cantidad y el tipo de gestos a reconocer.

Las personas están acostumbradas a utilizar dispositivos de hardware específicos como por ejemplo controles remotos, teclados, mouses o joysticks, para controlar distintos aparatos electrónicos. Incluso, algunos celulares pueden utilizarse como control remoto de electrodomésticos tales como televisores, aires acondiciones, etc. Más allá de la amplia gama de opciones, tener la posibilidad de manejar los dispositivos con el sólo movimiento de las manos, sin necesitar estos objetos, representa un avance tecnológico importante y significativo para muchas personas. Aunque ya existen diversos estudios sobre este tema, el problema no está totalmente resuelto.

Resolver la problemática de traducir los gestos a instrucciones ejecutables por un aparato electrónico resulta muy útil en diversos aspectos, no solo para eliminar hardware en algunos tipos de dispositivos sino también para personas con movilidad reducida que podrían operar distintos artefactos sin necesidad de desplazarse.

Dado que el reconocimiento de gestos hechos con las manos es un tema de mucha importancia y conociendo que existen varios trabajos al respecto, la motivación principal de esta tesina es lograr realizar el reconocimiento de una manera simple y con elementos de bajo costo. Por ello la captura del gesto hecho con las manos se realizará utilizando

una cámara de vídeo convencional y la señal correspondiente al comando identificado será transmitida de manera infrarroja.

Objetivo

El objetivo de esta tesina es desarrollar un algoritmo de segmentación de manos robusto y eficiente utilizando información de color, bordes y algunos operadores morfológicos. El paso inicial es la detección de las zonas de piel utilizando una red neuronal de arquitectura dinámica similar a la indicada en [36, 50]. Su entrenamiento requerirá la construcción de una base de datos de colores de piel ya que actualmente no se dispone de este tipo de información. La segmentación por color presenta inconvenientes ante cambios en la iluminación. Por tal motivo será necesario utilizar información adicional basada en distintas transformaciones así como reconocedores de bordes a fin de mejorar los resultados de la segmentación [51].

Se busca identificar gestos simples realizados con las manos para utilizarlos como comandos de entrada y ejecutar acciones en base a los mismos. Es decir, realizar un gesto, identificarlo y luego ejecutar la acción correspondiente al gesto. Para esto hay muchos aspectos a tener en cuenta, como por ejemplo cuáles son los gestos que interesa interpretar.

Una de las aplicaciones más famosas es utilizar gestos hechos con las manos para controlar las operaciones de un televisor [33] [6]. El objetivo es que en lugar de utilizar el control remoto habitual un usuario puede enviar un comando a un sistema reconocedor de gestos sólo con el movimiento de sus manos.

En esta tesina se propone utilizar el resultado de la segmentación de las manos unido al movimiento realizado durante el gesto como insumos para desarrollar además un reconocedor de gestos que será capaz de identificar un subconjunto básico de comandos permitiendo encender y apagar el televisor así como subir y bajar el volumen.

Organización del documento

Esta tesina aborda tanto el problema del reconocimiento de gestos como también su uso aplicado al control de dispositivos en este caso el control de un TV.

El documento está organizado de la siguiente forma:

- En el capítulo 1 se detallará que son y como se pueden clasificar los gestos hechos con las manos y además como se pueden interpretar mediante la interacción

humano-computador. También se detallarán distintos dispositivos de captura. En lo que se refiere a la cámara de vídeo, existen dispositivos que brindan información en 3D permitiendo captar no sólo el desplazamiento de los objetos sino la profundidad a la cual se encuentran con respecto el punto de observación [12]. Por ejemplo en [15] se utiliza un mapa de profundidad para ayudar a la segmentación. La elección de la cámara es una relación de compromiso entre el costo del equipamiento necesario para resolver el problema y la complejidad del algoritmo de segmentación a desarrollar. En esta tesina se ha decidido utilizar una cámara de vídeo convencional e identificar la zona en la cual se encuentran las manos a través de un reconocedor de colores de piel.

- En el capítulo 2 se introducirán los conceptos básicos necesarios para el procesamiento de imágenes digitales así como los sistemas de color existentes para representarlas. También se describen distintas estrategias de aplicación de filtros, control de ruido y detección de objetos.

Debe tenerse en cuenta que una mano en una imagen no tiene siempre las mismas características, ya que la palma y la cara de la mano son distintas en cuanto al color y los rasgos. Además, el tono de piel cambia de una persona a otra y las diversas condiciones de iluminación que pueden presentarse a la hora de capturar una imagen o un vídeo incrementan las dificultades del problema. Dado que existen diversos sistemas de color para representar una imagen digitalmente, es preciso identificar cual de todos ellos sería el más adecuado para resolver el problema de la detección de las manos. Entre los sistemas de representación más conocidos como RGB, HSI, HSL, HSV, TSL, CIE-Lab y YCbCr existen diferencias [13, 16, 40]. En esta tesina se analizarán y evaluarán las potenciales ventajas y desventajas de cada uno de ellos.

- En el capítulo 3 se describen las posibles técnicas utilizadas para la segmentación de manos, tanto las propuestas por otros autores, como también los productos ya existentes de las grandes empresas como son Kinect, Real Sense y LeapMotion.

El proceso de segmentación a partir del color generalmente se realiza a partir de umbrales predefinidos los cuales limitan la capacidad de ajuste del reconocedor [3, 35]. Obtener un algoritmo robusto para segmentar y detectar manos en una imagen o un vídeo tiene un gran potencial en el futuro ya que podría utilizarse para resolver múltiples problemas.

- El capítulo 4 se enfoca en describir que son las redes neuronales artificiales mencionando las más conocidas: el Perceptrón y el Multipercetrón con entrenamiento backpropagation. Esta última es una de las redes neuronales más famosas ampliamente utilizada como función de mapeo. Finalmente, en este capítulo se describe con detalle la red neuronal RCE ya que es la utilizada para implementar la solución propuesta en esta tesis.
- Luego en el capítulo 5 se describe en detalle como se completó el prototipo propuesto, tanto el software como el hardware que lo compone.

Luego de la segmentación debe procederse a representar la configuración y movimiento de la mano. Este paso nuevamente puede presentar distintos niveles de complejidad dependiendo de la diferencia entre gestos. En la literatura se pueden encontrar soluciones a problemas de reconocimiento de gestos dinámicos hechos con las manos sumamente complejos como el reconocimiento del lenguaje de señas Argentino [30] hasta la identificación por parte de un dispositivo de un conjunto acotado de señas muy distintas [21]. Esta etapa requiere del desarrollo de un clasificador robusto y es aquí donde las redes neuronales pueden aportar una solución; en especial las redes neuronales competitivas como la descripta en el capítulo 4.

- Finalmente en el capítulo 6 se detallan las conclusiones obtenidas de este trabajo y se exponen algunas líneas de trabajo futuras.

RECONOCIMIENTO DE GESTOS

1.1 Introducción

Los gestos fueron los primeros medios de comunicación que existieron en la antigüedad. Tienen como ventaja permitir la comunicación entre personas que hablan diferentes idiomas. Existen gestos universales cuyo significado puede ser interpretado por cualquier persona. Por ello el reconocimiento de gestos hechos con las manos tiene un papel muy importante a la hora de interpretarlos mediante un computador. Un software capaz de interpretar gestos hechos con las manos podría ser utilizado en cualquier parte del mundo sin necesitar ninguna modificación, lo cual no es un detalle menor. Por ello, la problemática de reconocer gestos y poder interpretarlos por un computador es un tema importante para muchos autores. Se han intentado varias maneras de procesar una señal generada por un gesto corporal/humano. Para esto pueden usarse diferentes técnicas, por ejemplo utilizar una cámara digital con un sensor de profundidad como es el caso de Kinect de Microsoft.

Las manos y los brazos, son las partes más móviles del cuerpo y por lo tanto ofrecen la capacidad de utilizarse para la comunicación no verbal. La forma más común es usarlas para señalar o acompañar la comunicación verbal. Por otro lado, también es la forma de comunicación más primitiva que existe, y naturalmente conocida, ya que lo primero que los humanos aprenden, antes de decir cualquier palabra, es a señalar o expresarse con gestos haciéndose entender sin expresar palabras en ningún idioma. La expresión gestual es más fácil de aprender que leer o escribir y además puede ser comprendida por

personas que hablan distintos idiomas.

1.2 Qué son los gestos

Un gesto es la manifestación de una expresión corporal la cual tiene un significado conocido por lo que puede ser interpretado por las distintas personas. Existe una amplia variedad de gestos posibles realizados por las personas. Para poder interpretar todas esas posibilidades se debe saber que todas las personas conocen esa misma señal y es interpretada de la misma manera, es decir, por ejemplo el gesto de saludar con la mano es conocido por la mayoría de las personas aunque hablen diferentes idiomas. Otro gesto universal es el apretón de manos. Son gestos sociales que están aceptados en las diferentes culturas. Pero existen otros en las diferentes culturas que tienen significados diferentes, como por ejemplo:

- Formar un círculo con el índice y el pulgar: significa Todo bien en Estados Unidos, Te mataré en Túnez y dinero en Japón.
- Palmas hacia arriba y dedos separados significa "Dámelo" en Estados Unidos y Reino Unido, Eres un tonto en Chile y es un gesto obsceno en Grecia.
- El Pulgar hacia arriba significa Sí, hagámoslo en Estados Unidos, Reino Unido y la mayor parte de Europa pero en el mundo árabe, partes de Italia y de Grecia es una Seña vulgar de desaprobación.
- El Puño cerrado en alto es el Signo de victoria en Reino Unido, Europa y es un gesto grosero o vulgar en Líbano y Pakistán.

Por esto cuando se habla de gestos se encuentran una gran cantidad de significados posibles en los diferentes idiomas y culturas.

Otra forma de utilizar los gestos para la comunicación es el lenguaje de señas, el cual es un herramienta fundamental para las personas que no tienen la capacidad de expresarse verbalmente, donde se comunican la través de señas hechas con las manos y brazos como así también para las personas que no tienen la capacidad de oír. Pero este tipo de lenguaje también depende del idioma que interpreta cada persona.

Como resumen de lo antes dicho, puede decirse que los gestos realizados con los brazos y manos son naturalmente conocidos por los seres humanos, y resultan una forma de comunicación naturalmente conocida.

1.3 Clasificación de los Gestos

Existen distintos trabajos de investigación referidos a la clasificación de los gestos humanos [39]. Los gestos tienen diferentes definiciones en diferentes contextos, y por lo tanto su clasificación basada en la definición del gesto podría variar en diferentes áreas. En el área de la lingüística y la ciencia cognitiva, el gesto es definido por McNeill en 1992 [23] como: "Los movimientos espontáneos de las manos y los brazos que se ven cuando la gente habla." Según esta definición, el gesto no incluye el movimiento que no acompaña al habla, mientras que en el contexto de las interfaces humano-computador, los gestos pueden ser realizados independientemente del habla y clasificados de forma independiente.

Los gestos son, sin duda alguna, uno de los aspectos más interesantes del comportamiento no verbal, y por supuesto lo más frecuentemente investigado dentro de este tema. El principal objetivo de los estudios planteados acerca de los gestos es establecer una relación entre éstos y los estados emotivos, atribuirles un significado o analizar sus funciones en relación a la comunicación verbal [29].

1.3.1 Clasificación de Efrón

Uno de los estudios más conocidos, sobre el significado de los gestos, es el trabajo de David Efrón en 1941 [7]; fue uno de los primeros en realizar un estudio exhaustivo de los gestos que acompañan al habla. En 1941 fue quien comenzó a estudiar los aspectos lingüísticos de los gestos de los judíos, orientales e italianos inmigrantes que residían en Nueva York. Él quería comparar como se involucraban las diferentes razas y culturas en el lenguaje corporal. Para ello analizó el comportamiento gestual, y observó que si bien los comportamientos verbales eran pronunciados de forma distinta en la primera generación, estos se iban unificando en las generaciones descendientes, esto es, cuanto más había asimilado un individuo las pautas gestuales autóctonas, exhibía menos gestos específicos de su grupo de origen. Efrón además concluyó que si un individuo se expone simultáneamente y durante un tiempo a la influencia de varios grupos, diferentes en sus gestos, adoptará y combinará ciertos comportamientos gestuales de todos ellos.

Efrón propuso dos categorías principales de gestos, según si el gesto tenía un significado independiente del discurso (objetivo), o junto con el discurso (lógico-discursivo).

Los gestos **objetivos** son aquellos que su significado no depende del discurso. Este tipo de gestos son identificados por la connotación que tienen independientemente del

contenido del discurso. De ellos existen las siguientes sub-categorías:

- Deíctico: Estos gestos están representados por el significado visual del objeto. Por ejemplo apuntar.
- Fisiográfico: Estos gestos representan, ya sea la forma de un objeto, una relación espacial o una acción corporal.
- Simbólico o emblemático: Estos gestos representan emblemas como por ejemplo el signo de OK de Norte América. Estos emblemas son comúnmente conocidos y además son tomados del vocabulario.

La lógica discursiva es la segunda categoría de los gestos definida por Efrom. Estos gestos tienen relación con el discurso de la persona, es decir, que acompañan el habla. No hacen referencia a un objeto sino que es una especie de representación gestual. Los gestos lógico-discursivos son una forma de representación gestual, que no hace referencia a un objeto, sino del curso del propio proceso ideacional. Por lo tanto, representan el proceso del pensamiento, en lugar de un objeto físico. En ella existen dos subcategorías:

- Bastones: Los bastones son gestos rítmicos que son similares al bastón de un conductor, y se utilizan para "vencer el ritmo de la locomoción mental".
- Ideográfico: El gesto ideográfico es un gesto que traza o dibuja en el aire los caminos y las direcciones del patrón de pensamiento. Efron da el siguiente ejemplo: un hablante que sacude su brazo en el aire entre las ubicaciones de dos tareas mentalmente imaginadas, y luego lo detiene en una de las ubicaciones como su conclusión.

Si bien la clasificación de Efron es demasiado abstracta para las aplicaciones actuales, la interacción humano-computador e interacción gestual con computadoras, abrió el camino para más estudios y su trabajo es considerado como una clasificación históricamente significativa por la comunidad.

Efrón, Ekman y Friesen [46] [47] han dado un fuerte impulso a la investigación en el campo de la gestualidad. Estos autores establecieron cinco categorías de señales no verbales, que aunque se refieren a los movimientos de todas las partes del cuerpo, definen especialmente los gestos de las manos. En su tipología distinguen entre emblemas, ilustradores, reguladores, señales de afecto y adaptadores, y señalan que estas categorías no poseen un carácter de exclusividad, de tal manera que un gesto no está incluido necesariamente en una sola de las categorías, pudiendo pertenecer a más de una.

1.3.2 Clasificación de Kendon

Luego de Efrón continuaron estudiándose los gestos. Uno de los trabajos más importantes fué el de Kendom [2]. Kendom definió el término *gesticulación* para lo que normalmente conocemos como gesto. Se trata de los gestos que acompañan la comunicación verbal, que ilustran el contenido del mensaje o su entonación. Kendom demostró que pueden encontrarse dependencias con el habla. De los más a los menos formales se encuentran: la gesticulación, gestos del lenguaje, pantomimas, emblemas y lenguaje de señas.

En los trabajos más recientes, se habla de gesticulación como la manera naturalmente ordinaria que la gente utiliza para la comunicación durante las conversaciones, especialmente cuando se dan descripciones, y las "interfaces de gesticulación" (interfaces de gesto naturales) son las interfaces del gesto y del habla donde los gestos acompañan el discurso para una interacción más natural. De acuerdo a esta definición, los gestos se basan en el análisis computacional y reconocimiento de patrones de movimientos de manos y no en un mapeo de gestos predefinido.

La diferencia entre pantomimas y emblemas en la clasificación de Kendon es que las pantomimas no son un conjunto predefinido de símbolos del vocabulario comúnmente conocido. Un emblema conocido es por ejemplo el signo de OK Norte Americano con el dedo indice y el pulgar formando un circulo.

La clasificación basada en la *formalidad* de Kendon, es muy importante ya que en la formalidad del mismo está involucrada en la capacidad de reconocer los gestos en los sistemas inteligentes. Las gesticulaciones tienen menor formalidad, mientras que los emblemas y los signos son más formales. Cuanto menos formal sea un gesto, más esfuerzo se requiere para reconocerlo y analizarlo con precisión, debido a la mayor ambigüedad e incertidumbre. A medida que aumenta la formalidad de los gestos, y se vuelven menos naturales, los usuarios tienen menos libertad para realizar gestos arbitrarios, lo que viola las promesas de investigación de interacción humano-computador que tienen como objetivo facilitar la interacción del usuario con las computadoras. Por ello estas investigaciones y los estudios de inteligencia artificial se enfocan en mejorar la flexibilidad de interacciones entre las interfaces y la gesticulación.

1.3.3 Clasificación de McNeill

Siguiendo a Kendon, McNeill en 1992 impulsó un punto de inflexión entre los estudios previos y las clasificaciones recientes sobre los gestos, donde se estudian los gestos independientemente del habla [23]. Estas investigaciones fueron continuadas por

Wexelbat en 1995 [43] y Quek en 2002 [9]. McNeill es uno de los autores más relevantes que habla del uso comunicacional de los gestos, quien plantea una clasificación conocida, que surge de estudiar los gestos de las personas.

La primera clasificación de McNeill, continuando con los primeros estudios de Efron, incluyen cuatro tipos: deícticos, icónicos, ilustrativos y metafóricos.

- Deícticos son aquellos movimientos que se utilizan para apuntar, identificar una entidad en particular. Estos son señalamientos, los que se pueden usar para señalar tanto objetos concretos como cosas abstractas.
- Icónicos son aquellos que representan una idea concreta, es decir, que el gesto tiene una relación estrecha con el contenido semántico del habla. Se puede decir que los gestos icónicos son "gestos de lo concreto", estos representan objetos mostrando sus características, es decir, pueden ser interpretadas a simple vista, por ejemplo, una mano elevándose representa algo o alguien subiendo. También pueden complementar el discurso haciendo movimientos que junto con la narración se interprete el objeto.
- Metafóricos representan una idea abstracta y son similares a los icónicos ya que representan imágenes, pero de un concepto abstracto, es decir, representan gráficamente la imagen de lo que se quiere expresar pero son una abstracción.
- Ilustrativos son aquellos que se sincronizan con el ritmo del discurso. Es decir, son movimientos simples y rápidos que acompañan el habla.

1.4 Clasificaciones más recientes

Las clasificaciones de Efron y McNeill están basadas en el discurso, y están limitadas en su aplicación en los sistemas interactivos. Los gestos que se realizan independientemente del habla humana necesitan ser clasificados de acuerdo a otros enfoques. A medida que la tecnología evoluciona, aumenta la necesidad de utilizar gestos naturales junto con herramientas de reconocimiento capaces de interpretar esta forma de comunicación.

1.4.1 Wexelblat

Wexelbat en 1998 [44] abordó la necesidad de diseñar e implementar sistemas capaces de reconocer con precisión gestos naturales, no solo gestos planteados o discretos. El estuvo

de acuerdo con los sistemas que sólo eran capaces de reconocer gestos pre planeados, artificiales o discretos. Creía que si los usuarios de un sistema tenían que realizar un gesto fijo para realizar una acción en dicho sistema, eso sería lo mismo que apretar una tecla.

Wexelblat considera que la interacción gestual natural es el único modo útil de interactuar con los sistemas informáticos: Uno de los puntos principales de los modos gestuales de operación es su naturalidad. Si se quita esa ventaja, es difícil ver por qué el usuario se beneficia de una interfaz gestual en absoluto.

1.4.2 Karam

Reviendo el estudio de Quek de 2002 [9] y otros trabajos relacionados, Karam en 2005 definió una taxonomía completa de las interacciones basadas en gestos. Karam consideró cuatro atributos principales para el clasificador: estilo de gesto, dominio de aplicación, tecnología de entrada y respuestas del sistema. Esto implica que si se considera un dominio de aplicación particular, entonces se están restringiendo dentro de ese dominio específico en términos de gestos de entrada / salida.

Las taxonomías de Karam son:

Estilo de gestos

En este caso Karam los categoriza en las siguientes clases:

- Deíctico: Esta clase es para gestos involucrados con apuntar para establecer la identidad / espacial ubicación de un objeto.
- Semáforos: Esta clase establece un sistema de signos y pistas que se pueden emplear para demostrar un significado. Los semáforos pueden ser dinámicos o estáticos. Los semáforos pueden ser expresados a través de la mano, los dedos, los brazos, la cabeza u otros objetos y dispositivos electrónicos, ratón.
- Gesticulación: Se considera como una de las formas más naturales de hacer gestos. Este tipo de gestos es comúnmente multimodal, consistente en movimientos de la mano en combinación por ejemplo. A diferencia de los semáforos, las gesticulaciones no se planifican ni enseñan.
- Manipulación: Esta clase establece una estrecha relación entre los movimientos de las manos y el objeto manipulado. Esta categoría se clasifica además en términos de Grado de libertad, como gestos bidimensionales o 3D.

- Gestos de lenguaje de señas: Los gestos de esta clase se basan en signos lingüísticos. A pesar de que son comunicativos, se diferencian de gesticulations ya que están pre-grabados

Entrada

Esta clase se basa en el tipo de dispositivo de entrada de la interacción. La clasificación de los gestos basada en el dispositivo de entrada se divide en dos categorías: Perceptual (como la visión por computadora o el reconocimiento de audio), que no requiere contacto físico y no perceptivo (como mouse y lápiz óptico, pantallas táctiles o interacción superficial) que requiere el tacto físico.

Dominio de aplicación

Esta dimensión se basa en el dominio al que se aplican los gestos, incluyendo: Realidad Aumentada y Virtual, Escritorio y Tablet PC, CSCW4, Pantallas 3D, Computación Ubicua y Entornos Inteligentes, Interfaces Móviles, Juegos, Telemática (TIC), Tecnología Adaptativa, Interfaces de Comunicación (estilo humano-humano de interacciones hombre-computadora).

Salida o respuesta del sistema

En esta categoría, los gestos se clasifican en función de la respuesta del sistema y del resultado real que conducen, incluidas las respuestas visuales y de audio además de las respuestas de comando de la CPU. Finalmente, cabe señalar que, aunque el trabajo de Karam y otros es una taxonomía amplia, carece de dimensiones específicas con la capacidad de captar rasgos principales de los gestos, como sus características físicas. Por lo tanto, no puede ser utilizado como un marco útil para que diseñadores e investigadores definan gestos.

1.5 Taxonomía de los gestos

Existen varias maneras de clasificar los gestos hechos con las manos, puede ser en base a las características observables o en las características de interpretación.

En la primera categoría los gestos se pueden dividir en dinámicos y gestos estáticos.

Se definen como gestos estáticos a aquellos en los que el usuario mantiene una posición o configuración, es decir, los gestos estáticos son aquellos en los que la posición

de la mano no cambia en un período de tiempo. Estos gestos se centran básicamente en las posiciones que pueden tomar los dedos.

Por otro lado, en los gestos dinámicos la posición de la mano cambia continuamente durante un período de tiempo, incluso en los gestos dinámicos se combinan movimientos de alguna parte del cuerpo con una o más posiciones de las manos.

Los gestos dinámicos generalmente tienen tres fases de movimiento: la preparación, el movimiento y la retracción. El mensaje en un gesto dinámico se encuentra principalmente en la secuencia temporal de la fase del movimiento. Los gestos dinámicos dependen tanto de las trayectorias, como de las orientaciones y flexiones de los dedos. Para ello es necesario muchas veces establecer un principio y un fin del gesto. En algunos casos se requiere poder interpretar tanto gestos estáticos como dinámicos, es así por ejemplo el caso del lenguaje de señas.

En el caso de la segunda categoría, los gestos se clasifican en función del significado interpretado. Por ejemplo, los símbolos son las clases típicas para describir esta clase de gestos. Los símbolos (también denominados como gestos autónomos) son gestos que se pueden sustituir por palabras habladas (por ejemplo, mostrando los pulgares arriba expresando que está todo bien). Los ilustradores son gestos usados para modelar las palabras habladas (por ejemplo, dando indicaciones al apuntar). La interacción y la comunicación entre el interlocutor y el oyente también son gestos conocidos.

Existen distintos tipos de gestos derivados de las partes del cuerpo con las que se expresan:

- Gestos hechos con las manos y brazos: se reconocen posiciones de las manos. Se utilizan en la lengua de señas y aplicaciones de entretenimiento.
- Gestos hechos con la cabeza y la cara: algunos ejemplos de estos gestos son asentir o negar sacudiendo la cabeza, guiñar un ojo, levantar las cejas, mover los ojos o la boca, etc.
- Gestos hechos con el cuerpo: pueden ser movimientos envolventes o hechos con todo el cuerpo como: movimientos de baile, reconocimiento de movimientos para rehabilitación médica, movimientos de entrenamiento deportivo, interacciones entre dos personas.

Entre estos distintos tipos de gestos pueden encontrarse variaciones dependiendo de su significado y de la interpretación del mismo. Además puede variar la forma de interpretación que se le da según características que dependen de la información que se obtiene del contexto. Los mismos pueden contener por ejemplo:

- Información espacial: relacionado con el lugar donde ocurre.
- Información temporal: relacionada con el camino que toma.
- Información simbólica: relacionada con el significado.
- Información afectiva: relacionada con la calidad emocional con la que se ejecuta.

Estas características de los gestos son fundamentales a la hora de intentar interpretarlos automáticamente mediante un computador. Las diferentes variaciones de los mismos dificultan las posibilidades de interpretar una gran cantidad de expresiones.

Las expresiones de la cara requieren que se identifiquen ciertas características como puntos de referencia tales como la nariz, la boca y/o los ojos. Estas marcas naturales permiten detallar el rostro de una persona y facilitan el proceso de detección y reconocimiento. En el caso de las manos, esta tarea no es tan simple ya que las manos no tienen características particulares como la cara. La fisonomía de la mano, no contiene detalles que permitan asegurar que el objeto detectado es realmente una mano. Al cambiar la posición, forma y/o tamaño, complejiza la solución para poder generar un algoritmo de reconocimiento.

En las imágenes se localizan estas partes para poder identificar un rostro. Es importante tener en cuenta la ubicación, intensidad y movimientos que pueden darse para poder identificar estas características en varios cuadros de vídeo. Es decir a la hora de reconocer gestos dinámicos, es probable que en una secuencia de frames consecutivos la ubicación de la boca no varíe considerablemente en el frame siguiente. Si bien esto parece simple no es aplicable a la identificación de gestos hechos con las manos ya que la mano no posee características tan particulares como el rostro. La fisonomía de la mano no tiene características específicas que se puedan identificar siempre. Esto se exemplifica en la Figura 1.1 donde si bien se observan dos manos, no es posible definir características que sean realmente descriptivas para las dos.

1.6 Clasificación de gestos en HCI

Los gestos pueden originarse de cualquier movimiento corporal o estado pero comúnmente se originan en la cara o en las manos. Los enfoques actuales en este campo incluyen el reconocimiento de la emoción en el rostro y de los gestos hechos con las manos. Los usuarios pueden usar gestos simples para controlar o interactuar con dispositivos sin tocarlos físicamente. Muchos enfoques se han hecho utilizando cámaras y



Figura 1.1: a) Mano cerrada; b) Mano abierta

algoritmos de visión por computadora para interpretar el lenguaje de señas. Sin embargo, la identificación y el reconocimiento de la postura, movimiento, y comportamientos humanos forman parte de los aspectos a resolver mediante las técnicas del reconocimiento del gesto. El reconocimiento de gestos puede ser visto como una manera de que las computadoras empiecen a entender el lenguaje corporal humano, construyendo así un puente más rico entre máquinas y seres humanos.

El reconocimiento de gestos permite a los seres humanos comunicarse con las máquinas (HCI) e interactuar naturalmente sin ningún dispositivo mecánico. Usando el concepto de reconocimiento de gestos, es posible señalar un dedo en la pantalla del ordenador para que el cursor se mueva en consecuencia. Esto podría hacer que los dispositivos de entrada convencionales como mouse, teclados e incluso pantallas táctiles sean redundantes. (Figura 1.2).

HCI (interacción hombre-computadora) es el estudio de cómo las personas interactúan con las computadoras y hasta qué punto las computadoras están o no desarrolladas para una interacción exitosa con seres humanos. Un número significativo de grandes corporaciones e instituciones académicas estudian HCI. Históricamente y con algunas excepciones, los desarrolladores de sistemas informáticos no han prestado mucha atención a la facilidad de uso de la computadora. Muchos usuarios de computadoras aún hoy argumentarían que los fabricantes de computadoras todavía no están prestando suficiente atención a hacer sus productos "fáciles de usar". Sin embargo, los desarrolladores de sistemas informáticos podrían argumentar que las computadoras son productos extremadamente complejos para diseñar y fabricar y que la demanda de servicios que las computadoras pueden proporcionar siempre ha superado la demanda de facilidad de uso.

Un factor importante de HCI es que diferentes usuarios forman diferentes concepciones o modelos mentales sobre sus interacciones y tienen diferentes maneras de

aprender y mantener el conocimiento y las habilidades (diferentes "estilos cognitivos" como, por ejemplo, "cerebro izquierdo" y "cerebro derecho" del ser humano). Además, las diferencias culturales y nacionales juegan un papel importante. Otra consideración al estudiar o diseñar HCI es que la tecnología de interfaz de usuario cambia rápidamente, ofreciendo nuevas posibilidades de interacción a las que los hallazgos de investigación anteriores podrían no ser aplicables. Por último, las preferencias de los usuarios cambian a medida que gradualmente dominan nuevas interfaces.

En el nivel más simple, se pueden desarrollar interfaces de gesto eficaces que responden a gestos naturales, especialmente al movimiento dinámico de la mano. Un ejemplo temprano es el Theramin, un instrumento musical electrónico de los años 20. Este responde a la posición de la mano utilizando dos sensores de proximidad, uno vertical y otro horizontal. La proximidad al sensor vertical controla el tono de la música, a la horizontal, la sonoridad. Lo asombroso es que la música se puede hacer con control ortogonal de las dos dimensiones principales, utilizando un sistema que no proporciona puntos de referencia fijos, como trastes o retroalimentación mecánica. Las manos trabajan de formas extremadamente sutiles para articular los pasos en lo que en realidad es un espacio de control continuo. El Theramin es exitoso porque hay una asignación directa del movimiento de la mano a la retroalimentación continua, lo que permite al usuario construir rápidamente un modelo mental de cómo utilizar el dispositivo [32].

Tipos de gestos

En otra clasificación considerada en las interfaces de computadora, se distinguen dos tipos de gestos acuerdo al tipo de acciones activadas por ellos. Dentro de esta clasificación se dividen en gestos en línea (online) y gestos fuera de línea (offline).

- Gestos fuera de línea: se procesan después de la interacción del usuario con el objeto, es decir, los gestos sin conexión se procesan después de que la interacción esté terminada. Estos generalmente transmiten la ocurrencia de un significado específico. Un ejemplo es el gesto para activar un menú.
- Gestos en línea: se procesan en el momento que se hace la interacción o realización. Se los conoce también como gestos de manipulación directa, que se utilizan por ejemplo para escalar o rotar un objeto tangible.

Dado que en esta tesina, se buscará reconocer gestos hechos con las manos para control de dispositivos, el enfoque será de gestos en línea.

1.6.1 Clasificación para su reconocimiento

En la interacción Humano-Computador es necesario diferenciar los distintos tipos de gestos para poder reconocerlos, se propone la siguiente clasificación según los aspectos de los gestos que impactan en los métodos de reconocimiento. [26].

- **Dinámicos y Estáticos:** Los gestos **dinámicos** son aquellos que están formados por secuencias de movimientos de una o más partes del cuerpo. Estos gestos involucran movimiento continuo en el tiempo por lo que necesitan reconocer el principio y el fin. Los gestos **estáticos** son aquellos que están compuestos por poses o configuraciones de posiciones del cuerpo.
- **Partes del cuerpo:** Para los gestos pueden involucrarse una única parte del cuerpo o varias. En el caso de los gestos dinámicos, pueden ser realizados por varias partes del cuerpo y en ese caso se debe tener en cuenta la sincronización ya que algunas partes del cuerpo pueden tardar distintos tiempos en realizar el movimiento o empezar y terminar en momentos diferentes. En los gestos estáticos se debe tener en cuenta el tiempo que tarda el cuerpo en alcanzar la posición deseada.
- **Faciales y corporales:** Los gestos **faciales** involucran el movimiento de los ojos, las cejas y los labios. Los gestos corporales involucran partes del cuerpo en el movimiento.
- **2D y 3D:** Los gestos realizados por las personas son tridimensionales pero para el reconocimiento se utilizan habitualmente sólo dos dimensiones evitando incluir la profundidad. Esto permite operar con técnicas más simples.
- **Unimodal y Multimodal:** Los sistemas unimodales son aquellos que tienen una única fuente de información para realizar el reconocimiento. Los sistemas multimodales pueden conformarse por varios tipos de sensores diferentes.
- **Dependientes o independientes del usuario:** Existen dos tipos de reconocimiento según su dependencia con el usuario. Los sistemas independientes del usuario contienen un conjunto de gestos definidos y los mismos son utilizados por todos los usuarios. Por el contrario los sistemas dependientes del usuario crean modelos de gestos para cada usuario en particular.

Los dependientes del usuario tienen como ventaja una mejor tasa de reconocimiento pero tienen como desventaja que deben ser entrenados por cada usuario que lo



Figura 1.2: Reconocimiento de Gestos

quiera utilizar. Los independientes del usuario tienen como ventaja que cualquier persona los puede utilizar en cualquier contexto.

La tecnología de reconocimiento de gestos ha sido considerada como una tecnología muy importante, ya que mejora los tiempos de uso en la mayoría de los dispositivos. El reconocimiento de gestos se puede realizar con técnicas de visión por computadora y procesamiento de imágenes. Las principales áreas de aplicación del reconocimiento de gestos en el escenario actual son: Sector automotriz, Sector de la electrónica de consumo, Sector de tránsito, Sector del juegos, para desbloquear smartphones.

1.6.2 Interfaces Touchless

Interfaz de usuario Touchless (TUI) es un tipo emergente de tecnología relacionada con el control de dispositivos a través de gestos. Interfaz de usuario Touchless es el proceso de comandar la computadora utilizando el movimiento del cuerpo y gestos sin tocar un teclado, ratón o pantalla [10]. Por ejemplo, Kinect de Microsoft es una interfaz de juego sin contacto; sin embargo, productos como el Wii no se consideran completamente sin contacto porque están atados a los controladores.

La interfaz Touchless, se está haciendo muy popular, ya que proporcionan la capacidad de interactuar con los dispositivos sin tocarlos físicamente [8].

Existen una serie de dispositivos que utilizan este tipo de interfaces con tecnología touchless, tales como, teléfonos inteligentes, computadoras portátiles, juegos y televisión. Aunque la tecnología sin contacto se ve principalmente en el software de juego, el interés se está extendiendo ahora a otros campos, incluyendo las industrias automotriz y de la

salud. Próximamente, la tecnología touchless y el control gestual se implementarán en los autos en niveles más allá del reconocimiento de voz como puede verse en [4].

1.6.3 Dispositivos de entrada

La capacidad de rastrear los movimientos de una persona y determinar qué gestos pueden estar realizando se puede lograr a través de diversas herramientas. La interface Kinetic (KUI) es un tipo emergente de interfaz que permite a los usuarios interactuar con dispositivos informáticos a través del movimiento de objetos y cuerpos. Ejemplos de KUI incluyen interfaces tangibles y juegos sensibles al movimiento como Wii, Microsoft Kinect, y otros proyectos interactivos.

Aunque hay una gran cantidad de investigación realizada en reconocimiento de gestos basados en imagen / video, hay cierta variación dentro de las herramientas y entornos utilizados entre implementaciones.

- Guantes con cable (Figura 1.3.a). Estos pueden proporcionar entrada a la computadora sobre la posición y la rotación de las manos usando dispositivos de seguimiento magnéticos o iniciales. Además, algunos guantes pueden detectar la flexión de los dedos con un alto grado de precisión (5-10 grados), o incluso proporcionar realimentación haptica al usuario, que es una simulación del sentido del tacto. El primer dispositivo de tipo guante de rastreo manual comercialmente disponible fue el DataGlove [38], un dispositivo de tipo guante que podía detectar la posición de las manos, el movimiento y la flexión de los dedos. Esto utiliza cables de fibra óptica que se ejecutan en la parte posterior de la mano. Se crean pulsos de luz y cuando los dedos están doblados, la luz se filtra a través de pequeñas grietas y la pérdida se registra, dando una aproximación de la pose de la mano.
- Cámaras de profundidad(Figura 1.3.b). Utilizando cámaras especializadas como la luz estructurada o las cámaras de tiempo de vuelo, se puede generar un mapa de profundidad de lo que se está viendo a través de la cámara a corto plazo y usar estos datos para aproximar una representación 3D de lo que se está viendo. Estos pueden ser eficaces para la detección de gestos de mano debido a sus capacidades de corto alcance [48].
- Cámaras estéreo (Figura 1.3.c). Usando dos cámaras cuyas relaciones entre sí se conocen, una representación 3D puede ser aproximada por la salida de las cámaras. Para obtener las relaciones de las cámaras, se puede usar una referencia de

posicionamiento, como rayos lexianos o emisores de infrarrojos. En combinación con la medición directa del movimiento (6D-Vision), los gestos pueden ser detectados directamente [17].

- Controladores basados en gestos(Figura 1.3.d). Estos controladores actúan como una extensión del cuerpo de modo que cuando se realizan gestos, parte de su movimiento puede ser convenientemente capturado por el software. Un ejemplo de la captura de movimiento basada en el gesto emergente es a través del seguimiento esquelético de la mano, que se está desarrollando para la realidad virtual y las aplicaciones de realidad aumentada. Un ejemplo de esta tecnología es mostrado por las empresas de rastreo uSens y Gestigon, que permiten a los usuarios interactuar con su entorno sin controladores [1] [20].

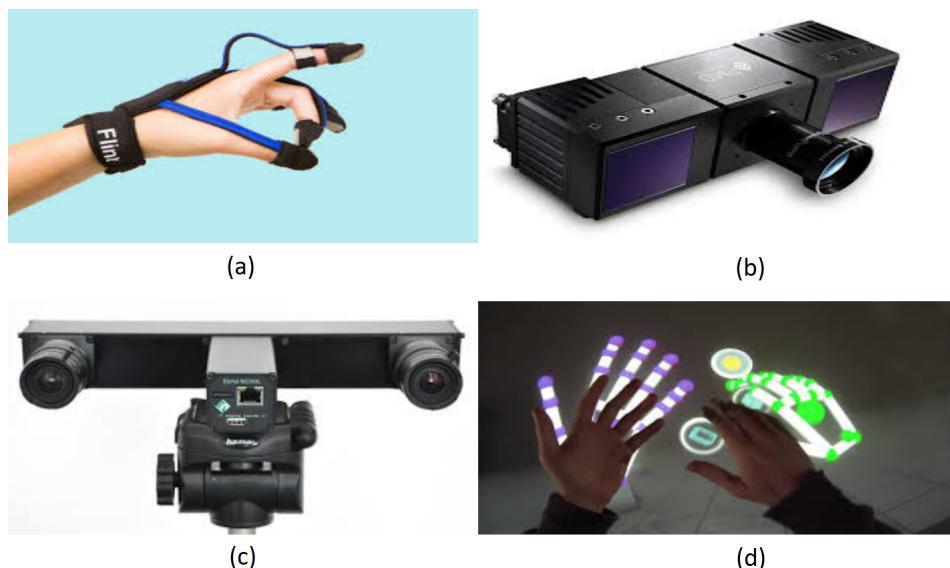


Figura 1.3: Dispositivos de entrada. (a) Guantes con cable. (b) Cámaras de profundidad. (c) Cámaras estéreo. (d) Controladores basados en gestos.

1.6.4 Algoritmos

Dependiendo del tipo de datos de entrada, el enfoque para interpretar un gesto podría hacerse de diferentes maneras. Sin embargo, la mayoría de las técnicas se basan en indicadores clave representados en un sistema de coordenadas 3D. Basándose en el movimiento relativo de estos, el gesto puede ser detectado con una alta precisión, dependiendo de la calidad de la entrada y el enfoque del algoritmo.

Para interpretar los movimientos del cuerpo, hay que clasificarlos según las propiedades comunes y el mensaje que los movimientos pueden expresar. Por ejemplo, en el lenguaje de señas cada gesto representa una palabra o frase. La taxonomía que parece muy apropiada para la Interacción Humano-Computador ha sido propuesta por Quek en [25]. En dicho trabajo se presentan varios sistemas gestuales interactivos para captar todo el espacio de los gestos: Manipulativo, Semaforico o Conversacional.

Algunas publicaciones diferencian dos enfoques en el reconocimiento de gestos: un modelo tridimensional y uno basado en la apariencia [42]. El primer método hace uso de información 3D de los elementos clave de las partes del cuerpo con el fin de obtener varios parámetros importantes, como la posición de la palma o ángulos de la articulación. El segundo modelo utiliza imágenes o videos para la interpretación directa.

Algoritmos basados en modelos 3D

El enfoque del modelo 3D puede utilizar modelos volumétricos o esqueléticos, o incluso una combinación de los dos. Los enfoques volumétricos han sido muy utilizados en la industria de la animación por computadora y con fines de visión por computador. Los modelos se crean generalmente a partir de superficies 3D complicadas, como NURBS o mallas poligonales.

B-splines racionales no uniformes o NURBS (en inglés non-uniform rational B-spline), es un modelo matemático muy utilizado en la computación gráfica para generar y representar curvas y superficies.

El inconveniente de este método es que posee un alto costo computacional, y los sistemas de análisis en tiempo real todavía están por desarrollarse. Por el momento, un enfoque más interesante sería mapear objetos primitivos simples a las partes más importantes del cuerpo de la persona (por ejemplo cilindros para los brazos y cuello, esfera para la cabeza) y analizar la forma en que éstos interactúan entre sí. Además, algunas estructuras abstractas y cilindros generalizados pueden ser aún más adecuadas para aproximar las partes del cuerpo. Lo interesante de este enfoque es que los parámetros para estos objetos son bastante simples. Para modelar mejor la relación entre éstos, se utilizan restricciones y jerarquías entre objetos.

Algoritmos basados en esqueleto

En lugar de utilizar el procesamiento intensivo de los modelos 3D y tratar con una gran cantidad de parámetros, se puede utilizar una versión simplificada de los parámetros de ángulo de la unión, junto con longitudes de segmento. Esto se conoce como una

representación esquelética del cuerpo, donde se calcula un esqueleto virtual de la persona y partes del cuerpo se asignan a ciertos segmentos. El análisis se realiza utilizando la posición y orientación de estos segmentos y la relación entre cada uno de ellos (por ejemplo, el ángulo entre las articulaciones y la posición u orientación relativa)

Como ventajas del uso de modelos esqueléticos puede decirse que los algoritmos son más rápidos porque sólo se analizan parámetros clave. Además es posible hacer coincidir patrones con una base de datos de plantillas. El uso de puntos clave permite que el programa de detección se enfoque en las partes significativas del cuerpo.

Modelos basados en apariencia

Muchos grupos de investigación en el reconocimiento de gestos utilizan métodos bastante complejos para reconocerlos, como la detección de dedos, el cálculo de los ángulos entre los dedos o la adaptación de los modelos en 3D. Para llevarlo a cabo, el sistema original se divide en varios subsistemas. Este enfoque tiene la desventaja de que un posible error se propaga a través de todo el proceso cuando el sistema padre toma una decisión equivocada. Las funciones utilizadas para el reconocimiento de gestos son:

- Color: brillo, modelos de color de piel.
- Textura: Filtros, gradientes.
- Forma: Formas Activas, Modelos de Contorno Activo.
- Movimiento: centroides, diferencias de imagen, flujo óptico.

Estos modelos no utilizan una representación espacial del cuerpo ya que derivan los parámetros directamente de las imágenes o vídeos usando una base de datos de plantillas. Algunos se basan en las plantillas 2D deformables de las partes humanas del cuerpo, articularmente las manos. Las plantillas deformables son conjuntos de puntos en el contorno de un objeto, utilizados como nodos de interpolación para la aproximación del contorno del objeto. Una de las funciones de interpolación más simples es la lineal, que realiza una forma promedio desde conjuntos de puntos, parámetros de variabilidad de puntos y deformadores externos. Estos modelos basados en plantillas se utilizan principalmente para el seguimiento manual, pero también pueden ser de utilidad para la clasificación simple de los gestos.

Un segundo enfoque en la detección de gestos utilizando modelos basados en apariencia utiliza secuencias de imágenes como plantillas gestuales. Los parámetros

para este método son las propias imágenes, o ciertas características derivadas de éstos. La mayoría de las veces, solo se usan una (monoscópica) o dos vistas (estereoscópicas).

1.7 Conclusión

En este capítulo se describió el concepto de gesto y la manera en que fueron definidos a lo largo de la historia por distintos investigadores como Efrón, Kendon y McNeill. A su vez se mencionó cómo evolucionaron esas definiciones con autores actuales como Wexelblat y Karam y cuales son las distintas clases de gestos existentes para poder introducir los criterios necesarios a la hora de interpretar los mismos.

Se introdujo el problema de reconocimiento de gestos, y cómo se realiza la interacción entre las personas y las computadoras. Se describió la manera de realizar la clasificación de los gestos de acuerdo a las formas, movimientos y significados. Además se describió cómo se los puede clasificar para realizar el reconocimiento dependiendo del dispositivo usado.

Este capítulo dejó en evidencia la gran cantidad de gestos diferentes que existen, y las técnicas que se utilizan para poder reconocerlos, teniendo en cuenta los distintos tipos dispositivos y formas de gestos posibles.

Esta tesina se enfocará en el estudio los gestos uni-modales, realizados con las manos. Para ello se empleará un técnica de dos dimensiones, usando una cámara digital convencional.

VISIÓN COMPUTACIONAL

2.1 Introducción

La adquisición de una imagen o escena física para convertirla en una imagen digital o estructura computacional está relacionada directamente con el hardware que se utiliza para ello. Existen distintos tipos de cámaras las cuales permiten obtener los datos de las mismas y abstraen totalmente de la electrónica necesaria para esto. El término de imagen digital incluye el procesamiento, compresión, almacenamiento y visualización de las imágenes.

La primera imagen digital fue producida en 1920, por el sistema de transmisión de imagen por cable Bartlane. Los inventores británicos, Harry G. Bartholomew y Maynard D McFarlane, desarrollaron este método. El proceso consistió en una serie de negativos en placas de zinc que se expusieron durante períodos variables de tiempo, produciendo así densidades variables. En 1957, Russell A. Kirsch produjo un dispositivo que generaba datos digitales que podían almacenarse en una computadora, esto se usó como un escáner y tubo fotomultiplicador.

En la actualidad existen tanto cámaras que permiten tomar imágenes de dos dimensiones, como también cámaras infrarrojas y sensores de profundidad que son capaces de registrar las tres dimensiones.

Una parte central de esta tesina está compuesta por las imágenes digitales, capturadas con una cámara digital de dos dimensiones. En este capítulo se describen como se representan las imágenes digitales en los sistemas de color que existen.

En las siguientes secciones se detallan las distintas técnicas para realizar el procesamiento de imágenes que permiten mejorar y filtrar los pixeles de la misma.

2.2 Conceptos básicos

Una imagen digital es la representación de una imagen mediante una matriz bidimensional de números en las que se almacena la información píxel a píxel del color que representa cada punto. El valor numérico de cada píxel corresponde a la información necesaria para visualizar la imagen respecto de su color, intensidad y luminosidad. Esto va a depender del sistema de color elegido por el cual se interpreta el valor almacenado en dicho pixel. De esta manera las imágenes pueden ser procesadas y visualizadas de acuerdo al sistema de color con el cual se encuentre almacenada.

Una imagen digital puede ser definida como una función de dos variables $f(x, y)$ donde x e y son coordenadas espaciales y la amplitud de f en el punto se define como intensidad. El resultado de esto es una matriz que representa la imagen. De acuerdo a los valores que puede tomar f definirá como está representada la imagen. Si f puede tomar sólo los valores 0 y 1, la imagen se visualizará en blanco y negro. Si f toma valores entre 0 y 255 la imagen se representará utilizando una escala de grises. Para el caso de las imágenes en color existe una amplia variedad de espacios de color que pueden utilizarse para representarla.

El sistema visual humano puede distinguir cientos de miles de colores e intensidades diferentes, pero sólo alrededor de 100 tonos de gris. Por lo tanto, en una imagen, una gran cantidad de información puede estar contenida en el color, y esta información adicional puede usarse entonces para simplificar el análisis de la imagen, ya sea en lo que se refiere a la identificación de objetos y/o su segmentación basada en color.

2.3 Imágenes digitales

Para la investigación de esta tesina es importante el debate sobre qué sistema de color debe utilizarse o resulta más adecuado para identificar patrones de piel. Para poder determinar que sistema de color es adecuado es necesario conocer como esta compuesto cada uno de ellos y que características son importantes para poder identificar los colores de piel.

Hay diferentes manera de representar los colores numéricamente, los cuales se detallan a continuación:

2.3.1 RGB

Una imagen RGB (Rojo, Verde y azul - Red, Green, Blue) está definida por una estructura de $M \times N \times 3$ pixels, donde cada pixel está compuesto por 3 valores correspondientes a los colores rojo, azul y verde. Este modelo comprende la composición de los colores mediante la mezcla de intensidades de los colores primarios.

La representación de RGB es obtenida por las siguientes fórmulas:

$$(2.1) \quad r = \frac{R}{R + G + B}$$

$$(2.2) \quad g = \frac{G}{R + G + B}$$

$$(2.3) \quad b = \frac{B}{R + G + B}$$

donde la suma de los componentes ($r+g+b=1$) es igual a 1. El tercer componente (el color azul) no contiene información significativa por lo que podría omitirse, reduciendo la dimensionalidad espacial. Los componentes restantes se los llaman a menudo "colores puros", porque la dependencia del rojo y verde en el brillo del color RGB fuente es disminuida por la normalización. Una propiedad notable de esta representación es que para las superficies mate, al ignorar la luz ambiental, el RGB normalizado es invariante (bajo ciertas suposiciones) a los cambios superficiales respecto a la fuente de luz. Esto ayuda a que el modelo RGB sea uno de los más populares.

En la figura 2.1 se puede observar como están distribuidos los colores en sus 3 dimensiones.

RGB es uno de los sistemas mayormente utilizado para representar imágenes por dispositivos como cámaras de fotos y video, monitores etc.

2.3.2 Matiz, Saturación, Intensidad

HS (matiz y saturación, en inglés Hue, Saturation) es un espacio de color basado en los valores de estas componentes. Describe intuitivamente los valores de saturación y tono. Hue define el matiz dominante en el área, la saturación mide el colorido de un área en cuanto al brillo. La intensidad de iluminación están asociados a la luminosidad del color.

me parece que en la frase anterior los sustantivos no están bien usados

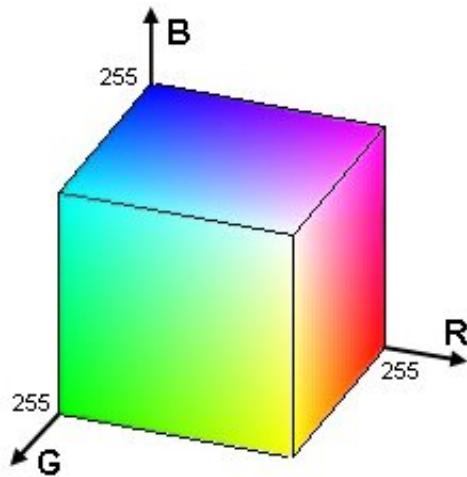


Figura 2.1: RGB

El espacio de color HS tiene distintas variantes como es el caso de HSI, HSV y HSL. HSI es representado por el matiz (H), saturación (S) e intensidad (I) el cual es similar a la forma que el ojo humano detecta los colores. HSL y HSV son las dos representaciones de coordenadas cilíndricas más comunes de puntos en un modelo de color RGB. Las dos representaciones reordenan la geometría de RGB en un intento de ser más intuitivo y perceptualmente relevante que la representación cartesiana (cubo). HSL significa matiz, saturación y luminosidad. HSV significa matiz, saturación y valor, también se lo denomina HSB (B para brillo).

La representación de HSV está dada por:

$$(2.4) \quad H = \arccos \left(\frac{(R - G) + (R - B)}{2\sqrt{(R - G)^2 + (R - B)(G - B)}} \right)$$

$$(2.5) \quad S = 1 - \left(\frac{3}{R + G + B} \right) + \min(R, G, B)$$

$$(2.6) \quad V = \frac{1}{3}(R + G + B)$$

En la figura 2.2 se puede ver el modelo de color HSI.

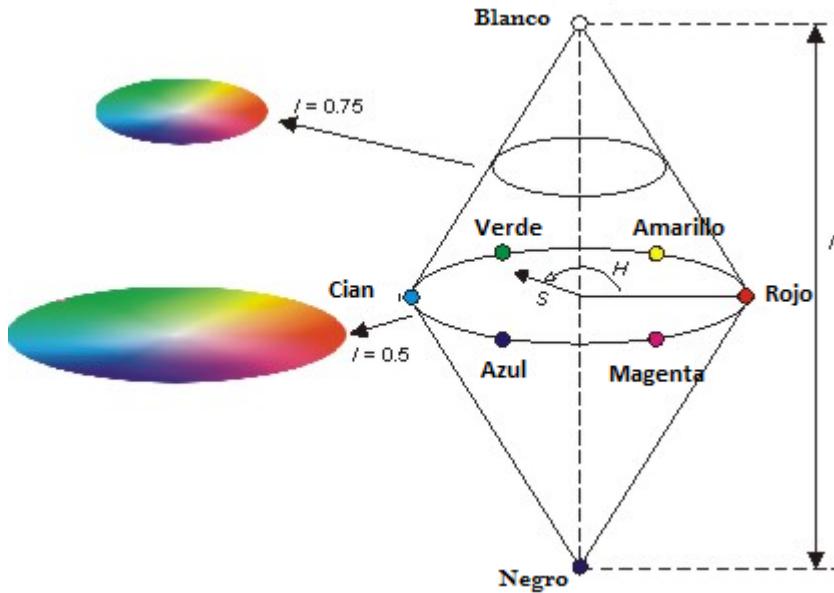


Figura 2.2: Modelo de color HSI

2.3.3 Matiz, Saturación, Luminosidad, TSL (Tint, Saturation, Lightness)

Otro espacio de color conocido es TSL el cual está compuesto por la luminosidad y el color. Es otra normalización de RGB un poco más intuitiva donde se centran la iluminación y saturación en un canal. La representación matemática de este sistema es la siguiente:

$$(2.7) \quad T = \begin{cases} \left(\frac{1}{2\pi}\right) \arctan\left(\frac{r'}{g'}\right) + \frac{1}{4} & g' > 0 \\ \left(\frac{1}{2\pi}\right) \arctan\left(\frac{r'}{g'}\right) + \frac{3}{4} & g' < 0 \\ 0 & g' = 0 \end{cases}$$

$$(2.8) \quad S = \sqrt{\frac{9}{5}(r'^2 + g'^2)}$$

$$(2.9) \quad L = 0.299R + 0.587G + 0.114B$$

donde:

$$(2.10) \quad r' = r - \frac{1}{3}$$

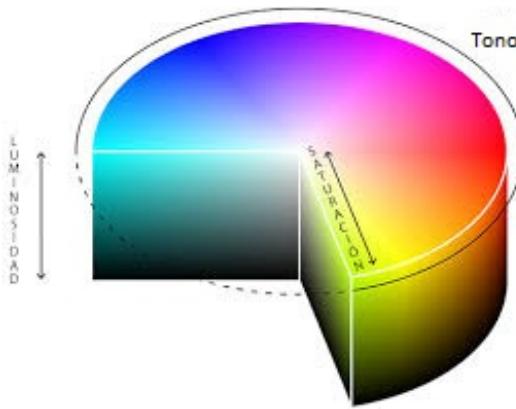


Figura 2.3: Modelo de color TSL

$$(2.11) \quad g' = g - \frac{1}{3}$$

$$(2.12) \quad r = \frac{R}{R + G + B}$$

$$(2.13) \quad g = \frac{G}{R + G + B}$$

Gráficamente la distribución de colores en TSL se pueden observar en la figura 2.3

2.3.4 YCrCb

YCrCb es un sistema de color que codifica de manera no lineal las señales RGB. Este espacio de color es representado por el valor de *luma* (luminosidad) calculado a partir de la suma ponderada de los valores RGB. Cr y Cb están formados por la resta de *luma* y los componentes rojo y azul de RGB. La representación matemática de este sistema es la siguiente:

$$(2.14) \quad Y = (0.2999R + 0.587G + 0.114B)$$

$$(2.15) \quad C_r = (R - Y)$$

$$(2.16) \quad C_b = (B - Y)$$

En la figura 2.4 se puede observar el modelo de este sistema de color.

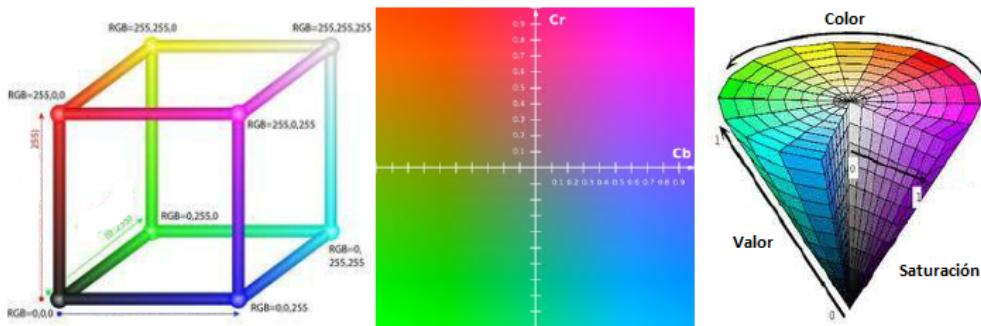


Figura 2.4: Modelo de color YCbCr

2.3.5 CIE-Lab

CIE-Lab es un espacio de color basado en el eje *L. Este es el eje de luminosidad (lightness) que va de 0 (negro) a 100 (blanco). También se utilizan otros dos ejes de coordenadas que son a* y b*, y representan la variación entre el color rojo y el verde, con el amarillo y azul, respectivamente. Aquellos casos en los que a* = b* = 0 son acromáticos; por eso el eje *L representa la escala acromática de grises que va de blanco a negro. Algunos espacios de color similares son CIE-Luv y Farnsworth UCS. La representación matemática de este sistema es la siguiente:

$$(2.17) \quad L = 116f\left(\frac{y}{y_n}\right) - 16$$

$$(2.18) \quad a = 500 \left[f\left(\frac{x}{x_n}\right) - f\left(\frac{y}{y_n}\right) \right]$$

$$(2.19) \quad b = 200 \left[f\left(\frac{y}{y_n}\right) - f\left(\frac{z}{z_n}\right) \right]$$

donde

$$(2.20) \quad f(t) = \begin{cases} t^{\frac{1}{3}} & s > 0.008856 \\ 7.787t + \frac{16}{116} & t < 0.008856 \end{cases}$$

En la figura 2.5 se puede observar el modelo de este sistema de color.

2.3.6 Comparación de los sistemas de color

Como se detalló anteriormente, existen distintas formas de representar las imágenes digitales de acuerdo al sistema de color elegido. Es necesario analizar cual es el sistema

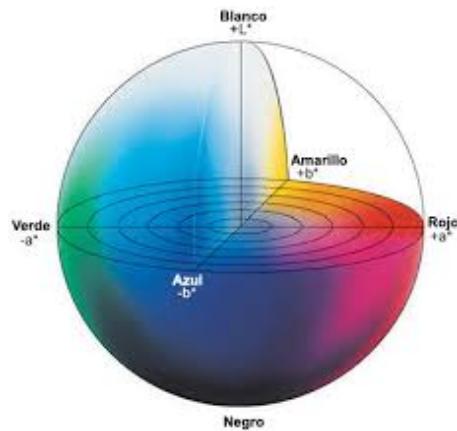


Figura 2.5: Modelo de color Cie-Lab

de color a utilizar que mejor se ajusta al problema a resolver. Abdesselam y Douglas [34] hicieron un estudio para comparar los distintos sistemas de color, aplicado a la segmentación de piel. Los autores utilizaron el clasificador Bayesiano con la técnica del histograma para analizar las diferentes representaciones de color.

Un clasificador Bayesiano es un clasificador probabilístico fundamentado en el teorema de Bayes y algunas hipótesis adicionales. El clasificador determina las pertenencias de un ejemplo a una clase aunque existan ausencias de algunas características. Es decir, la pertenencia o ausencia de una característica particular no está relacionada con la presencia o ausencia de cualquier otra, para una clase dada.

Según el estudio realizado por los autores, para el caso de este problema particular, el vector de características tiene una dimensión chica y tiene disponible un conjunto de datos grande. Por lo tanto, puede emplearse la técnica de histograma para la estimación de la pdf. También presentan una comparación con un clasificador Gaussiano y un multi-perceptrón.

La conclusión de dichos autores es que en cuanto a la representación de los colores, el estudio basado en el clasificador bayesiano muestra que la segmentación de la piel en píxeles no se ve afectada por la elección del espacio de color. Sin embargo, el rendimiento de la segmentación se degrada cuando sólo se utilizan canales de crominancia y hay variaciones significativas de rendimiento entre diferentes opciones. En términos de cuantificación de color, la más fina (un tamaño de histograma más grande) da mejores resultados de segmentación. Sin embargo, la estimación de la función densidad de probabilidad de color se puede hacer utilizando tamaños de histograma tan bajos como 64 bits por canal, siempre y cuando se use un conjunto de datos de entrenamiento grande

y representativo.

2.4 Procesamiento de Imágenes

El procesamiento de imágenes se centra en dos tareas: mejorar la información obtenida de las imágenes para la interpretación humana y realizar el procesamiento de datos para poder almacenar, transmitir y representar los datos de manera autónoma en un sistema.

Luego del análisis de los distintos sistemas de representación que existen para las imágenes, cabe mencionar que sólo con ello no es suficiente para poder obtener información sobre el contenido de la imagen. Es necesario analizar cada uno de los pixeles que componen dicha imagen para poder definir cuales son los objetos que están representados. Para lograr separar los objetos con sus bordes del fondo se deben aplicar diferentes técnicas que permitan diferenciar con claridad los contornos de las figuras que estén presentes. Para ello, es necesario tener en cuenta que las mismas, por ejemplo, pueden tener defectos. Los distintos tipos de cámaras pueden tener ruidos o capturar imágenes difusas. Esto dificulta su procesamiento. En estos casos es necesario aplicar sobre ellas funciones con las que se puedan aclarar o mejorar la definición de la imagen capturada.

El procesamiento digital de imágenes es el uso de algoritmos informáticos para operar sobre imágenes digitales. Como una subcategoría o campo de procesamiento de señales digitales, el procesamiento de imágenes digitales tiene muchas ventajas sobre el procesamiento de imágenes analógicas. Permite una gama mucho más amplia de algoritmos para ser aplicados a los datos de entrada y puede evitar problemas tales como la acumulación de ruido y distorsión de la señal durante el procesamiento. Dado que las imágenes se definen como mínimo en dos dimensiones, el procesamiento digital de imágenes puede ser modelado en forma de sistemas multidimensionales.

Los filtros digitales se utilizan para difuminar y afinar las imágenes digitales. El filtrado puede realizarse en el dominio espacial por convolución con núcleos diseñados específicamente (matriz de filtros), o en el dominio de frecuencia (Fourier) enmascarando regiones de frecuencia específicas. Los cuales se detallan a continuación.

2.4.1 Filtros de dominio de frecuencia

Los filtros de frecuencia procesan una imagen trabajando sobre el dominio de la frecuencia en la Transformada de Fourier de la imagen.

Transformada de Fourier

La transformada de Fourier, es una operación matemática utilizada para transformar señales entre el dominio del tiempo (o espacial) y el dominio de la frecuencia. En el caso de las imágenes digitales, las señales corresponden a los niveles de gris de la imagen o a la intensidad de grises de las diferentes filas o columnas de la matriz de una imagen. El eje temporal se reemplaza por los ejes x e y .

La transformada de Fourier se considera flexible para ser utilizada en la implementación de filtros o para la restauración de imágenes.

Definición: La transformada de Fourier hace corresponder a una función f con otra función g definida de la siguiente manera:

$$(2.21) \quad g(\xi) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} f(x)e^{-i\xi x} dx$$

Fourier indica que toda función periódica puede ser expresada como una suma infinita de senos y cosenos de infinitas frecuencias. El análisis que hace Fourier indica que a partir de una señal se puede determinar las frecuencias pero a costa de perder información temporal.

2.4.2 Filtros Espaciales

El término de dominio espacial en imágenes se refiere al plano de la imagen, y los métodos de esta categoría están basados directamente en la manipulación de los píxeles de una imagen.

Los filtros son uno de los métodos principales para realizar operaciones en las imágenes digitales. Pueden utilizarse para poder mejorar las imágenes de los ruidos mencionados, para eliminar las imperfecciones y facilitar la tarea de detección. Para esto se deben aplicar distintos tipos de filtros, los cuales permiten mejorar la calidad de las imágenes. Los filtros llamados espaciales, consisten en procesar cada pixel aplicando operaciones matemáticas que involucran a sus vecinos. Si los cálculos realizados sobre los píxeles de los vecindarios son lineales, la operación se denomina filtrado espacial lineal. Caso contrario el filtro sería no lineal. A continuación se explica cada uno en detalle.

Filtros Lineales

El concepto de filtrado lineal tiene sus raíces en el uso de la transformada de Fourier para el procesamiento de la señal en el dominio de la frecuencia. En este caso el interés está puesto en las operaciones de filtrado que se realizan directamente en los píxeles de una imagen. El uso del término **filtro espacial lineal** diferencia este tipo de procesos de los filtros de dominio de frecuencia.

Las operaciones lineales de interés en este caso consisten en multiplicar cada pixel del vecindario con el coeficiente correspondiente y sumar el resultado para obtener la respuesta en cada punto (x, y) . Los coeficientes están dispuestos como una matriz, llamados filtro o máscara.

Filtros NO Lineales

Una herramienta muy usada son los filtros NO lineales. Se utilizan para ordenar los valores de los píxeles vecinos de menor a mayor a partir de una lista ordenada. Los filtros espaciales **NO lineales** también se basan en operaciones sobre vecinos y en los mecanismos aplicados en los vecindarios de $M \times N$ deslizando el punto central a través de la imagen como en el caso de los filtros lineales. A diferencia de los filtros lineales, en este caso se aplican operaciones no-lineales involucrando a los píxeles del vecindario. Por ejemplo, dejar que la respuesta en cada punto central sea igual al valor máximo de píxel en su vecindario es una operación de filtrado no lineal. Otra diferencia es que el concepto de máscara no es frecuente en el procesamiento no lineal. Los filtros no lineales pueden ser de distintos tipos: máximos, mínimos o de mediana.

El **filtro de mediana** es utilizado para mejorar la calidad de la imagen y eliminar ruidos. Este filtro considera tanto a un pixel como a sus vecinos inmediatos para poder determinar si ese pixel es representativo o no del entorno. Simplemente determina el valor medio del pixel respecto de los valores que tienen sus vecinos. El valor medio se calcula ordenando los valores de los vecinos y se reemplaza el valor del pixel por el valor que es considerado en el medio de ese orden.

Este filtro es uno de los más simples, intuitivos y fáciles de implementar.

2.4.3 Morfología de la Imagen

En el lenguaje y la teoría de la morfología matemática se presentan dos puntos de vista de las imágenes binarias.

El término morfología se refiere a la descripción de las propiedades de forma y estructura de cualquier objeto. Las operaciones de la morfología matemática se definieron originalmente como operaciones en conjuntos, pero pronto se hizo evidente que también son útiles en las tareas de procesamiento del conjunto de puntos en el espacio bidimensional. Los conjuntos en morfología matemática representan objetos en la imagen.

La morfología matemática se utiliza para extraer algunas propiedades de la imagen que resultan útiles para su presentación y descripción; por ejemplo, contornos, esqueletos y cascós convexos. También se utilizan métodos morfológicos en el procesamiento preliminar y final de la imagen; por ejemplo, filtración morfológica, espesamiento o adelgazamiento.

2.4.3.1 Operaciones Morfológicas

La morfología matemática se basa en operaciones de teoría de conjuntos. En el caso de imágenes binarias, los conjuntos tratados son subconjuntos de Z^2 y en el de las imágenes en escala de grises, se trata de conjuntos de puntos con coordenadas en Z^3 . Las operaciones morfológicas simplifican imágenes y conservan las principales características de la forma de los objetos. Un sistema de operadores de este tipo y su composición, permite que las formas subyacentes sean identificadas y reconstruidas de forma óptima a partir de sus formas distorsionadas y ruidosas. Las operaciones de erosión y dilatación son elementales en la morfología matemática.

Erosión: Para la erosión, el elemento estructural pasa a través de todos los píxeles de la imagen. El píxel correspondiente a la referencia del elemento de estructura se activa si todo el elemento se corresponde con el área de primer plano (píxeles). Si todos los píxeles vecinos al pixel de estudio pertenecen al objeto, entonces el pixel de estudio también pertenece al objeto. Si alguno de los píxeles vecinos al pixel de estudio no pertenece al objeto entonces ese pixel de estudio tampoco pertenece al nuevo objeto.

Dilatación: La dilatación es una operación que "crece" o "engrosa" objetos en una imagen binaria. Si alguno de los píxeles vecinos al pixel en estudio pertenece al objeto entonces el pixel de estudio también pertenece al objeto. La dilatación de A y B es el conjunto de todos los desplazamientos z , tal que B y A se superponen en al menos un elemento.

Con las dilataciones y erosiones se modifica el volumen del objeto. Se decide combinar ambas operaciones, es decir, hacer tantas de una como de otra. Se define como apertura cuando: primero se hace una erosión y luego una dilatación (si había un hueco en el

interior de un objeto en la imagen A , se lo abre y si hay un objeto del tamaño de un pixel, se lo borra). En cambio, se define como Cierre cuando primero se hace una dilatación y luego una erosión (si había un hueco en el interior de un objeto en la imagen A desaparece. También se quitan rugosidades huecas de los bordes de los objetos). El objeto de estas operaciones es el de mantener el tamaño del objeto.

2.4.4 Ruidos

Las imágenes tomadas con cámaras digitales convencionales son propensas a tener imperfecciones, es decir, a tener pixeles que no tienen un color uniforme con el resto o que son producto de imperfecciones, movimientos, fallas del dispositivo con el que son tomadas. Por lo cual en muchos casos es necesario mejorar la calidad de las imágenes obtenidas para poder procesarlas con mayor exactitud.

Existen distintas técnicas para mejorar la calidad de estas imágenes, que permiten reconstruirla y realizar el procesamiento adecuado.

Una imagen digital tiene ruido cuando se puede identificar que hay una distorsión de la imagen que se quiso tomar realmente. Entonces se puede decir que aquella imagen que sufre una perturbación durante el proceso de adquisición, transmisión o almacenamiento tiene ruido. El ruido es también una degradación de la imagen que se manifiesta tomando valores distintos a los esperados.

Cuando obtenemos una imagen digital y se ve un poco distorsionada, se dice que la imagen tiene ruido. Según su definición, ruido es cualquier perturbación que sufre una señal en el proceso de adquisición y/o transmisión y/o almacenamiento. El ruido se modela usualmente como un proceso estocástico (modelo probabilístico). El ruido es un defecto de la información no deseado que contamina/degrada la imagen. Se manifestará generalmente en píxeles aislados que toman valores distintos de los reales.

2.4.4.1 Gaussiano

El ruido gaussiano está presente cuando el valor final del píxel es el real más una cierta cantidad de error y puede ser definido como una variable gaussiana que tiene distribución normal.

La función densidad de probabilidad p de una variable aleatoria Gaussiana z está dada por la siguiente ecuación:

$$(2.22) \quad p_G(z) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(z-\mu)^2}{2\sigma^2}}$$



Figura 2.6: Ejemplo Ruido gaussiano



Figura 2.7: Ruidos - Sal y pimienta

donde z representa el nivel de gris, μ el valor medio y σ la desviación estándar.

Este tipo de ruido produce pequeñas variaciones en la imagen que afectan la intensidad de los pixeles. Puede ser producido por defectos en el hardware utilizado para la captura de la imagen afectando a la imagen completa.

Se muestra en la figura 2.6 como se ve alterada un imagen con este tipo de ruido.

2.4.4.2 Sal y pimienta

Otro de los ruidos que pueden aparecer es conocido como sal y pimienta. Su presencia hace que distintos pixeles a través de la imagen tomen valores muy altos o muy bajos respecto de sus pixeles adyacentes provocando diferencias en las intensidades que podrían no tener que ver con la imagen real en si. Se muestra en la figura 2.7.

2.5 Segmentación de objetos

La segmentación de objetos en una imagen color implica particionar la imagen en diferentes regiones. Para determinar cuales son las regiones importantes que componen la imagen, se debe determinar cual es el problema a resolver.

Cada problema comprende distintos niveles de complejidad a la hora de determinar la posición de los objetos deseados en la matriz correspondiente a la imagen. Para ello se deben aplicar técnicas de correcciones de las imágenes, de filtrado y de restauración, las cuales serán descriptas a continuación.

La división de una imagen en estructuras significativas o segmentación, es a menudo un paso esencial en el análisis de imágenes, la representación de objetos, la visualización y muchas otras tareas de procesamiento de imágenes. Una gran variedad de métodos de segmentación se ha propuesto en las últimas décadas, y algunas categorías son necesarias para presentar correctamente los métodos.

En el caso de las imágenes digitales hay distintas elecciones posibles a la hora de procesarlas. Se puede elegir procesar los componentes de la imagen o procesar cada uno de los píxeles que la conforman. Para ello se necesitan conocer algunas técnicas que permitan determinar los objetos que hay en la misma como así también procesar los píxeles de acuerdo a sus intensidades convirtiéndolos en escalas de grises por ejemplo.

Otras de las consideraciones importantes, a la hora de detectar objetos, es el fondo. Este puede ser muy cambiante y a su vez puede interferir mucho a la hora de buscar algún elemento en la imagen. Se puede considerar como un fondo sencillo por ejemplo tener un color uniforme alrededor del objeto que se quiere identificar siempre y cuando el mismo no sea del mismo color del fondo. Ahora bien, esto no siempre es así, por lo general los fondos de las imágenes tienden a tener gran cantidad de objetos y mucha diversidad de colores. Si el objeto que se quiere tomar tiene un color que contrasta mucho con los colores que se pueden detectar en el fondo, la tarea resultará muy sencilla. A esto se le agregan muchos factores ya que también se debe tener en cuenta las formas y el tamaño de los mismos.

La segmentación de imágenes la divide en sus partes constituyentes hasta un nivel de subdivisión en el que se aíslen las regiones u objetos de interés. Los algoritmos de segmentación se basan en una de estas dos propiedades básicas de los valores del nivel de gris: discontinuidad o similitud entre los niveles de gris de píxeles vecinos.

Discontinuidad. Se divide la imagen basándose en cambios bruscos de nivel de gris:

- Detección de puntos aislados: Un punto aislado de una imagen tiene un tono de

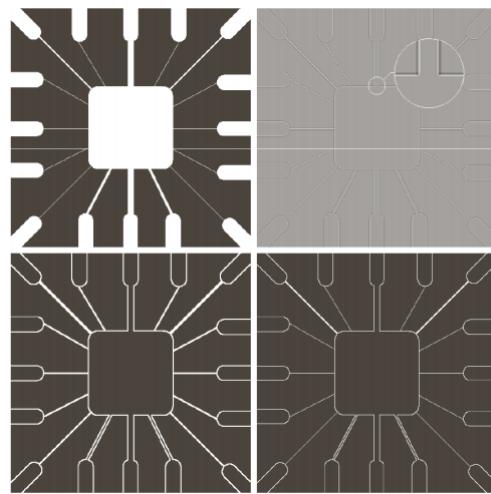


Figura 2.8: Máscaras de Laplaciano

gris que difiere significativamente de los tonos de gris de sus píxeles vecinos. Se dice que un píxel es un punto aislado si el resultado de aplicar la máscara sobre el píxel (en valor absoluto) es mayor o igual que un cierto valor umbral T , fijado por el decisor. Dicho valor depende de la aplicación que se esté realizando.

- Detección de líneas: Análogamente, para la detección de líneas de un píxel de ancho, se puede utilizar una máscara de Laplaciano. Se muestra en la figura 2.8.
- Detección de bordes: La idea que subyace en la mayor parte de las técnicas de detección de bordes es el cálculo de un operador local de derivación ya que un píxel pertenece a un borde si se produce un cambio brusco entre niveles de grises con sus vecinos. En general, no hay forma de conocer si los píxeles detectados como parte del borde son correctos o no de manera intuitiva. Es lo que se llama falso positivo (el detector devuelve un píxel cuando en realidad no pertenecía a ningún borde) y falso negativo (el detector no devuelve un píxel cuando en realidad pertenecía a un borde). Una manera posible de evaluar si un detector de bordes es bueno o no sería comparando el borde obtenido por el detector con el borde real de la imagen (para lo que, evidentemente, se necesita conocerlo de antemano). Existen otras aproximaciones que se basan en la "coherencia local". En este caso, no se compara con el borde real de la imagen, sino que se compara cada píxel detectado con sus vecinos.

Similitud. Se divide la imagen basándose en la búsqueda de zonas que tengan

valores similares, conforme a ciertos criterios prefijados:

- Crecimiento de región: Es un procedimiento que agrupa los píxeles o subregiones de la imagen en regiones mayores basándose en un criterio prefijado. Normalmente se empieza con unos puntos “semillas” para formar una determinada región y se van añadiendo aquellos píxeles vecinos que cumplan la propiedad especificada (por ejemplo, que estén en un rango de nivel de gris determinado). La propiedad considerada en el crecimiento de regiones debe tener en cuenta la información sobre conectividad o adyacencia de la imagen. Otro factor importante es la condición de parada.
- Umbralización: Un método básico para diferenciar un objeto del fondo de la imagen es mediante una simple binarización. A través del histograma se obtiene una gráfica donde se muestran el número de píxeles por cada nivel de gris que aparece en la imagen. Para binarizarla, se deberá elegir un valor adecuado (umbral) dentro de los niveles de grises, de tal forma que el histograma forme un valle en ese nivel. Todos los niveles de grises menores al umbral calculado se convertirán en negro y todos los mayores en blanco.

2.5.1 Descriptores

Para extraer características en una imagen es importante tener en cuenta sus diferentes componentes, uno de ellos por ejemplo es el fondo (background), es decir la región que se encuentra alrededor del límite del objeto identificado.

En visión por computadora, los descriptores visuales o descriptores de imágenes son descripciones de las características visuales de los contenidos en imágenes, videos o algoritmos o aplicaciones que producen tales descripciones. Describen características elementales como la forma, el color, la textura o el movimiento, entre otros.

Una región es un componente conectado, y el significado (también llamado borde o contorno) de una región es el conjunto de píxeles en la región que tienen uno o más vecinos que no están en la región. Los puntos que no están en un límite o región son puntos de fondo de llamada. Inicialmente sólo interesan las imágenes binarias, los puntos de límite son representados por los puntos 1s y el fondo por 0s.

Dada la definición del párrafo anterior, se dice que un límite es un conjunto de puntos conectados. Los puntos de un límite se ordenan si forman una secuencia en sentido horario o anti-horario. Se dice que un límite está conectado minuciosamente si cada uno

de sus puntos tiene exactamente dos vecinos de 1 valor que no son adyacentes a 4. Un punto interior se define como un punto cualquiera en una región, excepto en su límite.

Los descriptores son el primer paso para descubrir la conexión entre los píxeles contenidos en una imagen digital y lo que los humanos recuerdan después de haber observado una imagen o un grupo de imágenes después de algunos minutos.

Los descriptores visuales se dividen en dos grupos principales:

- Descriptores de información general: contienen descriptores de bajo nivel que describen el color, la forma, las regiones, las texturas y el movimiento.
- Descriptores específicos de información de dominio: proporcionan información sobre objetos y eventos en la escena. Un ejemplo concreto sería el reconocimiento facial.

2.6 Conclusión

En este capítulo se describió como están compuestas las imágenes digitales y cuales son las distintas formas de representarlas en los distintos espacios de color.

Se describieron en detalle funciones matemáticas aplicables a las imágenes digitales para realizar tanto el mejoramiento como el filtrado de las mismas para poder realizar una segmentación óptima.

Las técnicas descriptas en este capítulo permiten mejorar las imágenes obtenidas por la cámara web para poder tener mejor precisión a la hora de la segmentación.

La información de este capítulo está basada en el libro [10].

En los capítulos siguientes se utilizarán estos conceptos ya que se procesarán imágenes digitales para realizar el reconocimiento de gestos en los métodos propuestos de interacción humano-computador.

SEGMENTACIÓN DE MANOS

3.1 Introducción

El reconocimiento de gestos hechos con las manos es de gran importancia para la interacción hombre-máquina, debido a la gran variedad de aplicaciones en las que se puede utilizar, como realidad virtual, reconocimiento del lenguaje de señas, y vídeo juegos. A pesar de que existen numerosos trabajos realizados en este campo, las operaciones basadas en métodos de reconocimiento gestos hechos con las manos todavía están lejos de ser satisfactoria para aplicaciones de la vida real. Debido a la naturaleza de la detección óptica, la calidad de las imágenes capturadas es sensible a las condiciones de iluminación y a los fondos complejos, por lo tanto, los métodos basados en sensores ópticos suelen no ser demasiado robustos. Para sortear estos problemas las grandes empresas construyeron soluciones que implican la utilización de equipos de hardware más costosos.

En el capítulo anterior se introdujo en detalle como están compuestas las imágenes digitales. Es necesario tener en cuenta esta información para poder realizar la segmentación de manos en las mismas. Existen diferentes técnicas, las cuales serán descriptas en este capítulo.

Como se detalló en el capítulo 1, dado que el reconocimientos de gestos es una interfaz de comunicación hombre-máquina muy intuitiva, las grandes empresas desarrollaron productos como Kinect y Realsense los cuales de describen en este capítulo.

3.2 Comunicación hombre-maquina

A la hora de procesar la comunicación entre el hombre y las máquinas sin utilizar hardware específico, es necesario arbitrar los mecanismos para obtener la información para controlar a los dispositivos. Hay distintas maneras de introducir información a un computador sin usar los elementos convencionales como el teclado y el mouse, o una pantalla táctil, como por ejemplo los comandos por voz.

En los últimos años a crecido el interés sobre la interacción hombre-máquina, conocido como HCI (Human-Computer Interaction). Se busca mejorar el uso de los dispositivos electrónicos mediante mecanismos más amigables para los seres humanos. Dado que los gestos son un medio naturalmente conocido, por el cual las personas se comunican, es necesario desarrollar una técnica con la cual las computadoras personales puedan reconocer lo que se quiere comunicar. Poder integrar una interfaz de comunicación que pueda interpretar los gestos que hacen las personas reduce la brecha comunicacional que existe entre el lenguaje humano y la interpretación de un computador.

La elección de utilizar cámaras convencionales en esta tesina está dada por que los costos de éstas suelen ser muy inferiores a los costos de los otros dispositivos y además, tienen la posibilidad de utilizarse en cualquier lugar. Hoy en día cualquier celular tiene una cámara digital con lo cual esto hace que sea un elemento que puede ser accesible por cualquier persona. Esto refiere a la utilización de una cámara de dos dimensiones en la cual la única información que se puede obtener es la ubicación, es decir, las coordenadas x e y de la posición de la mano u objeto a detectar.

En lugar de utilizar solo una cámara digital convencional algunos autores utilizan otro tipo de sensores para capturar el gesto y el movimiento, con el propósito de mejorar la robustez de estos métodos. Por ejemplo una de las posibles técnicas es la utilización de guantes en las manos. El guante a diferencia de otros tipos de sensores ópticos, suelen ser más confiables y no se ven tan afectados por condiciones de iluminación o fondos complejos. Sin embargo, se requiere que el usuario utilice un objeto particular y además a veces requiere de calibración. El inconveniente de utilizar tanto este como otro tipo de marcadores es que puede obstaculizar la articulación del gesto de la mano.

Otra técnica utilizada son los sensores de profundidad, que permiten contar no solo con la posición sino también con la distancia que hay al sensor. Por su intermedio se puede obtener fácilmente cual es la posición de la mano y si la misma está más adelante que el resto del cuerpo. Este método también tiene sus desventajas, una de ellas tiene que ver con los costos asociados al hardware y en el caso de arquitecturas cerradas, el

3.3. PRODUCTOS UTILIZADOS PARA RECONOCER GESTOS

costo de las licencias para poder utilizarlo.

Otra de las opciones que refieren distintos autores, también relacionada a los marcadores, es la utilización de pulseras a partir de las cuales se puede determinar donde comenzar la segmentación.

El hecho de poder encontrar una manera sencilla de realizar la comunicación hombre-máquina no es una tarea fácil. Usar señales o gestos permiten que esa comunicación resulte más amigable para cualquier potencial usuario.

3.3 Productos utilizados para reconocer gestos

Como las manos son una de las partes del cuerpo más hábiles del ser humano, se cree que estas serían un dispositivo de entrada realmente eficiente y natural para la comunicación con las nuevas tecnologías. Aunque esto se cree muy importante, todavía hay múltiples dificultades a resolver para poder obtener un reconocimiento lo suficientemente robusto. Existen dos grandes formas de plantear esta problemática, una es desde la clasificación basada en la detección de objetos, y otra es por segmentación de color.

Para la detección de objetos basados en la clasificación, se requiere un gran número de comparaciones, lo que puede reducir drásticamente la velocidad de procesamiento del sistema. También se puede utilizar un esquema jerárquico de clasificadores que elimina falsos positivos en la parte superior y aplica clasificadores más precisos en el nivel inferior lo cual podría reducir considerablemente el tiempo de procesamiento. Otra manera similar, es construir un árbol de decisiones agrupando los datos en formas similares. Sin embargo, ambos métodos requieren una gran cantidad de datos de entrenamiento, entre otras.

La segmentación por color de piel es un método bastante común para agrupar partes de la mano en secuencias de imágenes, debido a que puede implementarse rápidamente por umbrales fijos o tablas de color. Aunque se han propuesto diversos métodos para segmentar partes de piel en imágenes, la segmentación por color sigue siendo una de las tareas más difíciles en el reconocimiento de gestos, y determina como será la calidad de la clasificación. Generalmente, es preciso considerar tres aspectos a la hora de segmentar la piel por color: representación de color, cuantificación de color y esquema de clasificación. Cuando se habla de representación de color está referido a seleccionar un espacio de color adecuado para detectar los pixeles que contienen colores de la piel. La cuantificación de color es para determinar el número de colores distintos para describir el color de la piel. En cuanto al esquema de clasificación, es un proceso para decidir las áreas de la piel en

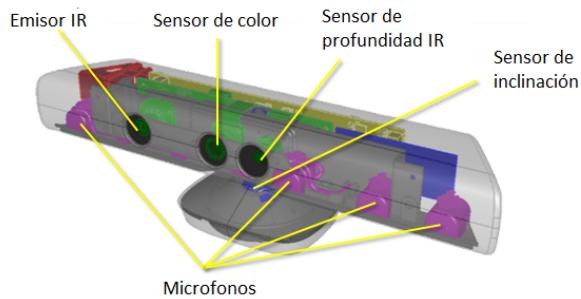


Figura 3.1: Kinect de Microsoft

las imágenes.

En esta sección se describen algunos de los dispositivos, existentes actualmente en el mercado, aplicables al reconocimiento de gestos hechos con las manos.

3.3.1 Kinect- Microsoft

Kinect es un dispositivo que cuenta con una cámara web, un sensor de profundidad y cuatro micrófonos con los cuales es capaz de reconocer movimiento, gestos, rostros y voz. Además posee un acelerómetro para determinar su orientación respecto del suelo. Este dispositivo fue desarrollado por Microsoft acompañado de su SDK que permite detectar personas y partes del cuerpo en tiempo real. Este software es propietario con lo cual se requiere una licencia para utilizarlo en el desarrollo de aplicaciones. La Figura 3.1 muestra una imagen de este dispositivo.

El sensor de profundidad está compuesto por un láser infrarrojo y una cámara infrarroja monocromática. Además contiene un sensor CMOS. La tecnología de detección utilizada es desarrollada por la empresa PrimeSense, la cual no da a conocer por completo las características del hardware utilizado. Aunque se sabe que está basado en el principio de luz estructurada; este principio es el utilizado por los scanners.

Mediante la cámara infrarroja y el sensor, se reconstruye la imagen en tres dimensiones, mediante triangulación. Dado que los patrones de puntos son aleatorios, la coincidencia entre la imagen infrarroja y los patrones detectados por el protector pueden compararse de manera directa entre los vecinos y utilizar correlación cruzada normalizada.

Entre la cámara infrarroja y el sensor de profundidad se obtiene una mapa de profundidad, como se muestra en la figura 3.2. Este mapa es el resultado de los valores obtenidos de profundidad en escala de grises, donde los píxeles de color negro son aquellos



Figura 3.2: Sensor de profundidad de Kinect

que están fuera del alcance del sensor y a medida que los píxeles son más blancos, la distancia al sensor es menor.

En algunos casos los valores de profundidad detectados por el sensor, no son correctos. Esto sucede cuando la calibración entre el proyector y la cámara infrarroja es incorrecta. Puede ser producido por calor o movimiento o un desvío en el infrarrojo. Para resolver este problema el equipo de Microsoft tuvo que desarrollar la posibilidad de recalibrar el sensor manualmente. Cuando se identifica que el dispositivo no está detectando los movimientos de manera adecuada hay que proceder a realizar la recalibración.

Este dispositivo permite de una manera fácil, determinar donde se encuentra posicionada la mano de una persona o cualquier parte de su cuerpo, con sus limitaciones. Como ya se ha mencionado, es un software propietario por lo que es necesario comprar la licencia para usar el SDK y además es necesario adquirir el hardware específico.

Esta es una solución bastante robusta que permite identificar las manos de una persona pero el costo que conlleva tanto el dispositivo como el desarrollo encarece la posibilidad de que sea una solución escalable y adquirible por todas las personas. Kinect permite detectar las manos y los gestos hechos con ellas pero no permite detectar donde están los dedos por lo que también dificulta la posibilidad de reconocer algunos tipos de señas.

3.3.2 RealSense- Intel

Intel también se involucró en la problemática del reconocimiento de gestos. En este caso a través del desarrollo del producto RealSense que permite interactuar mediante reconocimiento de gestos, facial y de voz. Al igual que en el caso de Kinect la cámara

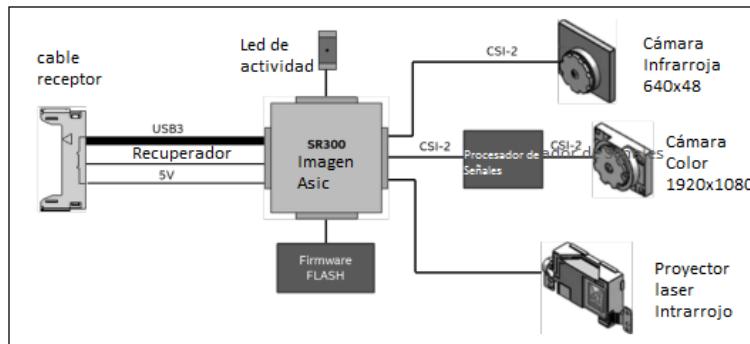


Figura 3.3: Sistema de imágenes 3D - Real Sense - Intel

utilizada para realizar el reconocimiento no es solo un dispositivo convencional de dos dimensiones, sino que también utiliza hardware específico. En este caso la cámara RealSense SR300 está compuesta por dos cámaras, una cámara color full HD, una cámara infrarroja VGA monocromática que mide la profundidad de los objetos y además un láser también infrarrojo que se utiliza para contemplar las dimensiones. También contiene dos micrófonos que se utilizan para el reconocimiento de voz. En la figura 3.3 se muestra en detalle la arquitectura del dispositivo RealSense.

El proyector infrarrojo y la cámara infrarroja forman una matriz de píxeles monocromáticos, que determinan una imagen de profundidad en dos dimensiones. Estos valores son procesados por el sistema de procesamiento específico(ASIC) de generación de imágenes para las tramas de vídeo con profundidad y/o infrarrojo que se transmiten al sistema cliente a través de USB3. La cámara de color consta de un sensor cromático y un procesador de señal que captura la imagen y procesa los valores de los píxeles cromáticos. Estos valores generan tramas de vídeo en color que se transmiten al ASIC de imagen y luego se transmiten al sistema cliente a través de USB3. La cámara de color puede funcionar independientemente de la cámara infrarroja o pueden funcionar de forma sincrónica para crear frames de vídeo de color, infrarrojos y/o profundidad.

Para generar el frame infrarrojo, el proyector ilumina la escena con un patrón blanco. Esto genera un reflejo que es capturado por la cámara infrarroja. Los píxeles capturados por la cámara son procesados por el ASIC de generación de imágenes para generar el frame monocromático. El ASIC convierte el frame recibido en una imagen 3D de la escena capturada. La figura 3.4 muestra el flujo de procesamiento.

A diferencia de Kinect de Microsoft, la tecnología Real Sense que produjo Intel es capaz de realizar el seguimiento tanto de manos como de los dedos de la mano. El SDK permite obtener la posición de la punta de los dedos como también del antebrazo. El SDK

3.3. PRODUCTOS UTILIZADOS PARA RECONOCER GESTOS

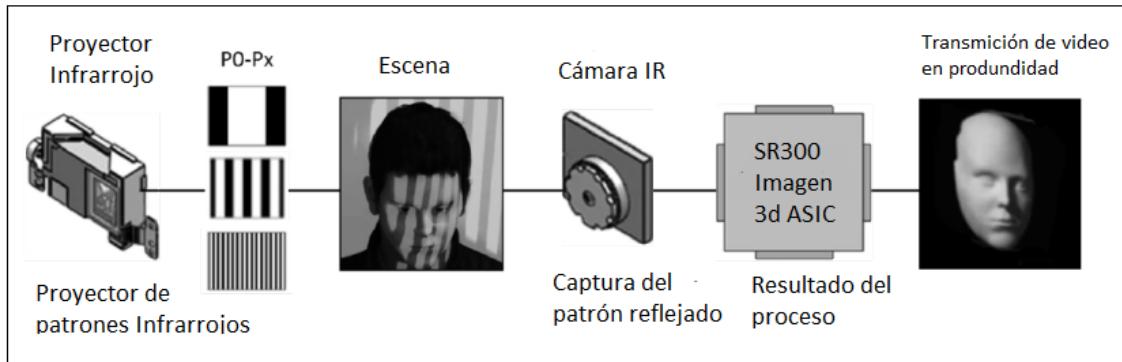


Figura 3.4: Flujo de captura de imagen en profundidad - Real Sense

de 2014 mejoró en cuanto a la detección de manos y permite obtener 22 puntos para el seguimiento de la mano. Para el reconocimiento de gestos, permite identificar 8 poses estáticas y 6 gestos dinámicos que actúan como una forma natural de interactuar con el software. Algunas de las poses estáticas que permite reconocer son: pulgares para arriba y el signo de la paz.

Intel permite mediante el SDK utilizar las funciones de la cámara Realsense para desarrollar las aplicaciones. Al igual que Kinect, el dispositivo Real Sense es una solución bastante robusta para conseguir la detección de la posición de la mano y el tracking de la misma. Como mejora sobre Kinect se puede observar que este dispositivo permite encontrar los dedos de la mano y ya contiene la detección de algunas posiciones y gestos. Pero también tiene el costo agregado del hardware y el SDK, lo cual encarece el desarrollo de las aplicaciones.

3.3.3 Leap-Motion

Leap Motion Inc. es una empresa estadounidense que fabrica y comercializa un dispositivo de hardware para computadoras que permite sensar movimientos de las manos y dedos, análoga a un ratón, pero no requiere contacto con la mano. En 2016, la compañía lanzó un nuevo software diseñado para el seguimiento manual en la realidad virtual. La tecnología para Leap Motion se desarrolló por primera vez en 2008.

Leap-Motion es un dispositivo de control mediante gestos con gran precisión en los movimientos de las manos. El controlador Leap Motion es un pequeño dispositivo periférico USB que está diseñado para colocarse en un escritorio físico, mirando hacia arriba. También se puede montar en un auricular de realidad virtual. Usando dos cámaras IR monocromáticas y tres LEDs infrarrojos, el dispositivo observa un área



Figura 3.5: Leap Motion

aproximadamente semiesférica, a una distancia de aproximadamente 1 metro. Los LEDs generan luz IR sin patrón y las cámaras generan casi 200 fotogramas por segundo de datos reflejados. Esto se envía a través de un cable USB al ordenador host, donde es analizado por el software Leap Motion usando "matemáticas complejas" de una manera que no ha sido revelada por la compañía, sintetizando datos de posición 3D comparando el marco 2D generado por las dos cámaras. En un estudio de 2013, la precisión promedio general del controlador se mostró a 0,7 milímetros. (Figura 3.5).

El área de observación más pequeña y la mayor resolución del dispositivo diferencian el producto del Kinect, que es más adecuado para el seguimiento de todo el cuerpo en un espacio del tamaño de una sala de estar. En una demostración a CNET-Networks [37], se mostró que el controlador realizaba tareas tales como navegar por un sitio web, usar gestos de pinch-to-zoom en mapas, dibujos de alta precisión y manipular complejas visualizaciones de datos 3D.

3.4 Experiencias en segmentación de manos

Dado que la problemática de reconocer gestos hechos con las manos, sin utilizar hardware especializado no es una tarea fácil de resolver, diversos autores apelaron a múltiples técnicas que permitan realizar esta tarea en tiempo real.

Una de las formas más simples para obtener este resultado es utilizando umbrales de color fijos. Estos umbrales se fijan en un rango adecuado para el color de la piel humana.

3.4. EXPERIENCIAS EN SEGMENTACIÓN DE MANOS

Ahora bien, es bastante complejo utilizar este tipo de técnicas, ya que la piel humana comprende un amplio conjunto de colores, siendo necesario aplicar rangos diferentes para distintos tonos de piel. Esto último dificulta bastante utilizar este tipo de técnicas para resolver problemas complejos.

También pueden utilizarse distintos tipos de marcadores, es decir figuras u objetos en las imágenes con colores particulares, con los cuales se pueden usar como punto de referencia para encontrar la mano o la mano marcada en si.

Tanto en el caso de Kinect como de Real Sense, se deja de lado la problemática de las imágenes a color, dado que la detección está centrada en el dimensionamiento de la profundidad percibida por los lasers infrarrojos.

3.4.1 Marcadores

Varios autores utilizan cámaras digitales convencionales, junto con algún tipo de marcador, el cual permite rápidamente encontrar el objeto buscado en la imagen.

Así es el caso por ejemplo de los guantes para detectar gestos con las manos o interpretar el lenguaje de señas [30] o por ejemplo [19] donde utilizan pulseras blancas en las muñecas para detectar donde comienza la mano. Estos marcadores aceleran la detección de los objetos, ya que, de acuerdo al color elegido del marcador y a las características del fondo, se puede definir la utilización un umbral de color con el cual determinar las coordenadas x e y del objeto en la imagen.

Utilizar este tipo de técnicas facilita la tarea de detectar los objetos en una imagen, de manera simple y además ayuda a resolver problemas más complejos, dejando de lado esta problemática que no es trivial de resolver.

Como mencionan los autores, al utilizar el marcador, sólo se necesita buscar el color determinado en la imagen, para luego procesarlas y en algunos casos realizar algunas verificaciones para determinar si la forma del objeto es la correcta. Luego se puede utilizar esta información para resolver problemas complejos. Esta información se utiliza por ejemplo, para reconocer configuraciones de la mano y poder interpretar el lenguaje de señas o poder utilizar la mano como dispositivo de entrada en una computadora personal.

También se utilizó este método para manejar un robot con instrucciones simples para direccionarlo y manejar su velocidad. Esta prueba fue realizada como proyecto de alumnos de la Facultad de Informática, y demostró que esta técnica también tiene problemas cuando los fondos son complejos y la iluminación cambia. Al tener diversidad de colores en el fondo, en algunos casos se puede perder el objeto en cuestión si la verificación de la forma no es lo suficientemente robusta. Para el problema de la

iluminación se utilizó una calibración manual que estabiliza los parámetros de acuerdo a la iluminación del ambiente actual.

3.4.2 Clasificación para lenguaje de señas

El trabajo [30] se enfoca en el problema de clasificación de configuraciones de manos. Está centrado en la extracción de características representativas de la mano y en el reconocimiento de dichas configuraciones utilizando una variante de red neuronal competitiva supervisada denominada ProbSom. Los autores de este trabajo generaron una base de datos de configuraciones de manos para el Lenguaje de Señas Argentino (LSA), junto con un modelo de procesamiento de las imágenes y clasificación de las configuraciones.

Los autores sortean el problema de la segmentación de la mano utilizando guantes de color y los sujetos utilizaron guantes de tela con colores fluorescentes en sus manos. Las imágenes son tomadas con sujetos que visten ropa negra, sobre un fondo blanco con iluminación controlada. Las consideraciones anteriores resuelven el problema de segmentación pero no en su totalidad, sin embargo es lo suficientemente eficaz para el reconocimiento de la posición de la mano, y deja de lado el problema del color de piel.

Este trabajo presenta una base de datos de configuraciones de Lengua de Señas Argentina creada con el propósito de producir un diccionario de LSA y entrenar un traductor automático de señas con configuraciones de mano utilizadas en distintas señas de dicho lenguaje. Los autores tuvieron en cuenta las diferentes posiciones y rotaciones en el plano perpendicular a la cámara, para generar mayor diversidad y realismo en la base de datos.

Se muestra en la figura 3.6 un ejemplo de las imágenes de este trabajo.

3.4.3 Reconocimiento del alfabeto

El proyecto de reconocimiento de gestos hechos con las manos usando la visión por computadora [19] construye una interfaz hombre-máquina utilizando una cámara de vídeo para interpretar el alfabeto de un solo idioma y el número de gestos. El sistema de cámara elegido en este trabajo comprendió una matriz 2D de píxeles RGB proporcionada a intervalos de tiempo regulares. Revisar el párrafo anterior

Para detectar la información de la silueta fue necesario diferenciar la piel de los píxeles de fondo. Se utilizaron otros marcadores para proporcionar información adicional sobre la mano. Es importante considerar que los píxeles del marcador también tendrán

3.4. EXPERIENCIAS EN SEGMENTACIÓN DE MANOS

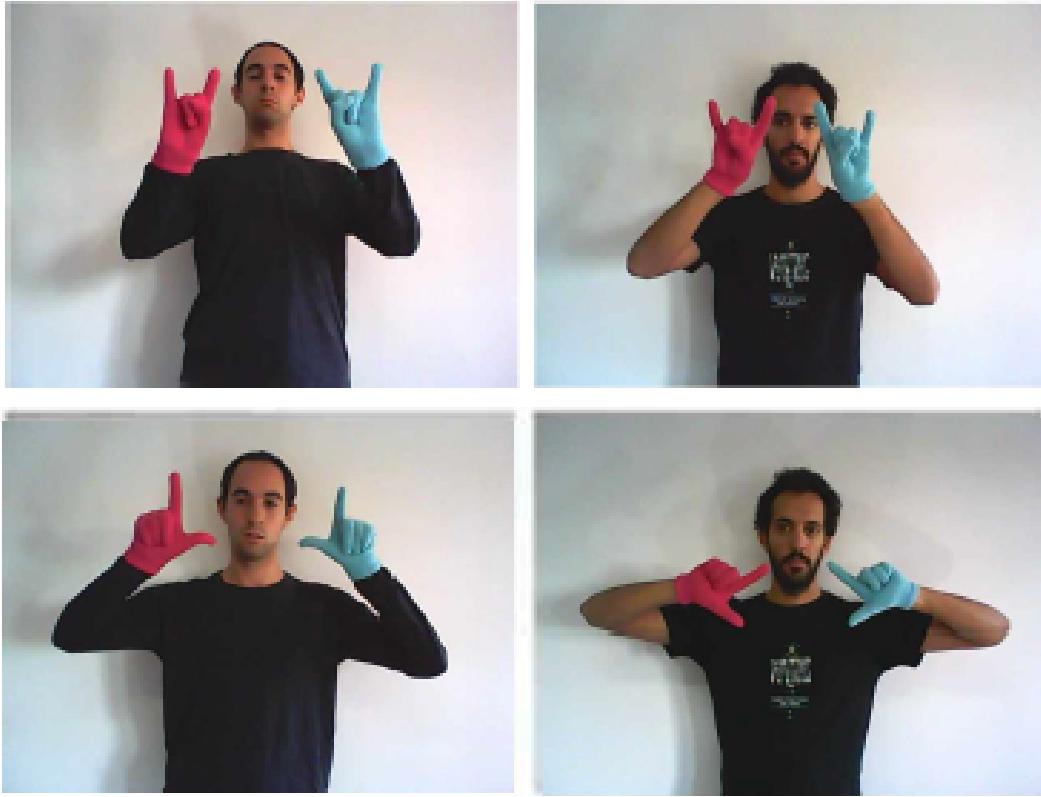


Figura 3.6: Imágenes no segmentadas de la base de datos LSA

que diferenciarse del fondo (y píxeles de la piel). Para que este proceso sea factible, es esencial que la configuración del hardware se elija cuidadosamente.

En el trabajo [18] se planteó hacer el reconocimiento de gestos mediante el método de aprendizaje. En este método, el gesto establecido para ser reconocido es enseñado al sistema de antemano. Cualquier gesto dado entonces es comparado con los gestos almacenados y una puntuación de coincidencia es calculada. El gesto de puntuación más alto puede mostrarse si su puntuación es mayor que un cierto umbral de calidad de concordancia. La ventaja de este sistema es que no requiere información previa sobre las condiciones de iluminación o la geometría de la mano para que el sistema funcione, ya que esta información es codificada en el sistema durante el entrenamiento. El sistema es rápido si el conjunto de gestos se mantiene pequeño. La desventaja con este sistema es que cada gesto necesita ser entrenado por lo menos una vez y para cualquier grado de exactitud, varias veces. El conjunto de gestos puede ser específico del usuario.

La imagen 3.7 muestra las áreas de calibración del color para la banda de muñeca (verde) y la piel (naranja). La calibración se realiza colocando la banda de muñeca bajo el

área de calibración verde y la mano debajo del área de calibración de color naranja (la imagen 3.8 muestra una mano parcialmente posicionada). El algoritmo de calibración lee los valores de color de ambas áreas y calcula los rangos actualizando repetidamente los valores máximo y mínimo RGB para cada píxel.



Figura 3.7: Calibración A

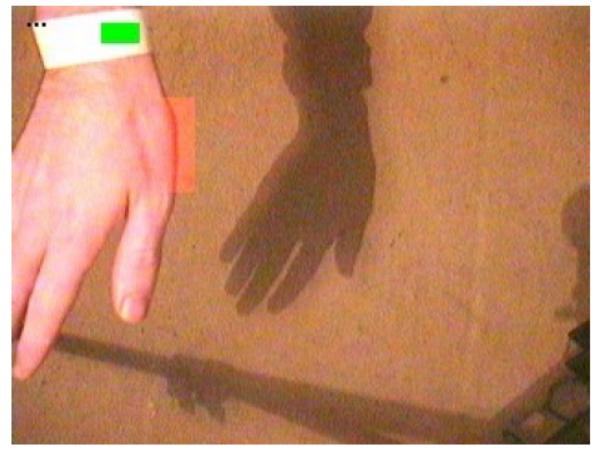


Figura 3.8: Calibración B

3.5 Conclusión

En este capítulo se describieron las técnicas utilizadas por los distintos autores para realizar la segmentación de manos. Se consideraron las técnicas desarrolladas por las grandes empresas para poder realizar el reconocimiento de gestos como son Microsoft e Intel con sus productos Kinect y RealSense respectivamente.

Esta soluciones son costosas y en algunos casos se encuentran basadas en código propietario por lo que se consideró el uso de cámaras web estándar para captar el gesto. En este capítulo se describieron distintas técnicas que pueden ser utilizadas en este caso.

Antes de pasar a detallar el método propuesto en esta tesina es preciso introducir los conceptos básicos de redes neuronales y su aplicación a problemas de clustering o agrupamiento ya que constituyen una parte central de la solución propuesta. El siguiente capítulo introduce este tema.

REDES NEURONALES

4.1 Introducción

Las redes neuronales biológicas han inspirado el diseño de redes neuronales artificiales. En la neurociencia, una red neuronal biológica es una serie de neuronas interconectadas cuya activación define una vía lineal reconocible. La interfaz a través de la cual las neuronas interactúan con sus vecinos generalmente consiste en varios terminales axónicos conectados a través de sinapsis a dendritas de otras neuronas.

En este capítulo se introducirá el concepto de redes neuronales artificiales, y además se describen las más comunes. Luego se describirá en detalle la red neuronal RCE (Restricted Coulomb Energy) la cual se utilizó para la implementación del prototipo que será descripto mas adelante.

4.2 Definición

Una red red neuronal artificial (RNA) es un paradigma de procesamiento de información basado en la forma en que los sistemas nerviosos biológicos procesan la información que busca simular el comportamiento del cerebro humano. Algunas características atractivas de las redes neuronales artificiales que las hacen superiores a otras técnicas de inteligencia artificial son:

CAPÍTULO 4. REDES NEURONALES

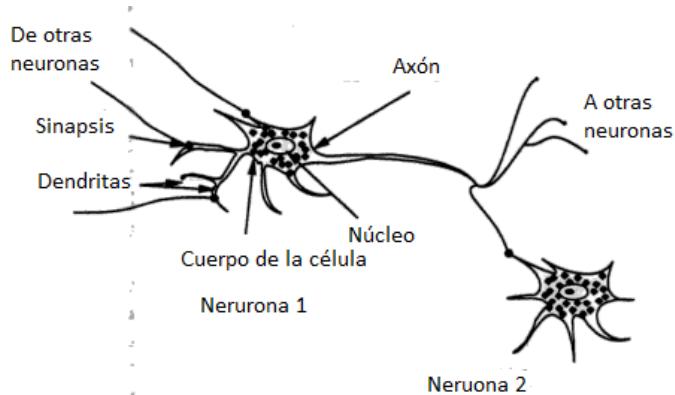


Figura 4.1: Neurona

- Robustez y tolerancia a fallos.
- Flexibilidad
- Capacidad de adaptarse a un diversidad de situaciones
- Gran capacidad de procesamiento

La arquitectura y el modo de funcionamiento de las RNA se basa en las redes neuronales biológicas.

Tal como puede verse en la figura 4.1, las neuronas consisten en un cuerpo celular llamado soma donde se encuentra el núcleo de la neurona y se conectan a través de fibras nerviosas llamadas dendritas. El estímulo que recibe la neurona ingresa por las dendritas y se acumula en el soma hasta que supera un cierto umbral; cuando esto ocurre el estímulo es propagado hacia otras neuronas a través de una única salida denominada axón, donde eventualmente se ramifican y se conectan con otras neuronas realizando nuevas sinapsis. Los extremos receptores de estas funciones en otras células pueden encontrarse tanto en las dendritas como en los propios cuerpos celulares. La transmisión de señales desde una célula a otra, es decir la sinapsis, es un proceso químico complejo, en el que son liberadas sustancias desde el emisor. El efecto de esto es el aumento o disminución del voltaje eléctrico dentro de la célula receptora. Si el potencial eléctrico alcanza el umbral, la actividad eléctrica genera impulsos cortos. Cuando esto sucede se dice que la célula se ha encendido. Estas señales eléctricas tienen una fuerza y duración fija que son enviadas hacia el axón. Generalmente la actividad eléctrica se da en el interior de las neuronas mientras que la actividad química se da en la sinapsis. Las dendritas sirven como receptores de señales entre las neuronas, mientras que el

propósito del axón es transmitir la actividad neuronal entre las celdas, dentro de la neurona o las fibras musculares. El tamaño del cuerpo celular de una neurona típica es aproximadamente entre 10 y 80 milímetros. [49]

4.2.1 Redes neuronales artificiales

Una red de neuronas artificial puede considerarse como un modelo simplificado de las redes neuronales biológicas. Las redes neuronales artificiales son modelos computacionales basados en el funcionamiento de conjuntos de unidades neuronales simples, simulando el comportamiento observado en los axones de las neuronas en los cerebros biológicos.

Una red neuronal se define como: la tupla ordenada (N, V, ω) , donde N y V son dos conjuntos y ω es un función. N es el conjunto de neuronas y V un conjunto $(i, j) | i, j \in N$ cuyos elementos se llaman conexiones entre neurona i y neurona j . La función $\omega : V \rightarrow R$ define los pesos, donde ω_{ij} , el peso de la conexión entre la neurona i y la neurona j , se acorta a ω_{ij} . Dependiendo del punto de vista es indefinido o 0 para las conexiones que no existen en la red.

Una red neuronal artificial se basa en una colección de unidades conectadas llamadas neuronas artificiales, (análogo a los axones en un cerebro biológico). Cada conexión (sinapsis) entre las neuronas puede transmitir una señal a otra neurona. La neurona receptora (postsináptica) puede procesar la/s señal/es y luego señalar a las neuronas conectadas a ella. Las neuronas pueden tener un estado, generalmente representado por números reales, típicamente entre 0 y 1. Las neuronas y sinapsis también pueden tener un peso que varía a medida que avanza el aprendizaje, lo que puede aumentar o disminuir la fuerza de la señal que envía hacia abajo. Además, pueden tener un umbral tal que sólo si la señal agregada está por debajo (o por encima) de ese nivel es la señal descendente enviada.

Normalmente, las neuronas se organizan en capas. Diferentes capas pueden realizar diferentes tipos de transformaciones en sus entradas. Las señales se desplazan desde la primera (entrada), hasta la última capa (salida), posiblemente después de recorrer las capas varias veces.

El objetivo original del enfoque de la red neural era resolver los problemas de la misma manera que lo haría un cerebro humano. Con el tiempo, la atención se enfocó en combinar capacidades mentales específicas, conduciendo a desviaciones de la biología como la re-propagación (backpropagation), o pasando la información en la dirección inversa y ajustando la red para reflejar esa información.

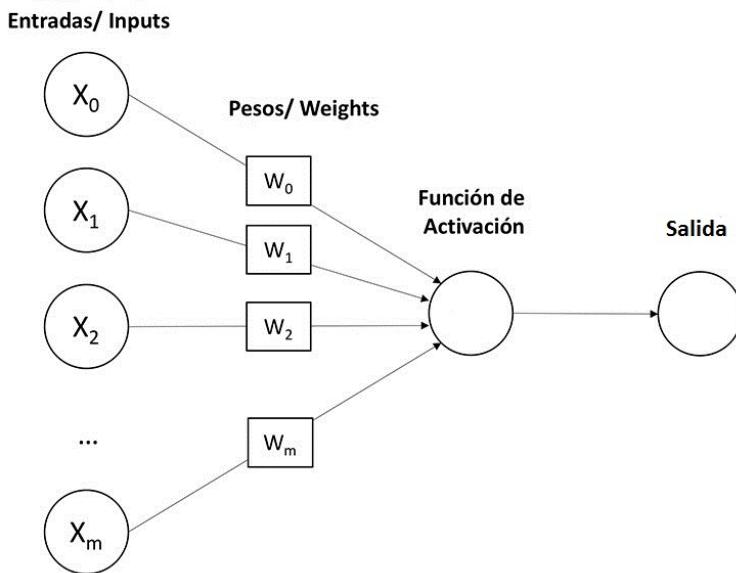


Figura 4.2: Arquitectura Perceptrón

Las redes neuronales se han utilizado en una variedad de tareas, incluyendo visión por computadora, reconocimiento de voz, traducción automática, filtrado de redes sociales, juegos de mesa y video juegos, diagnóstico médico y en muchos otros dominios.

Warren McCulloch y Walter Pitts [1] (1943) crearon un modelo computacional para redes neuronales basado en matemáticas y algoritmos llamados lógica umbral. Este modelo abrió el camino para la investigación de redes neuronales para dividir en dos enfoques. Un enfoque se centró en los procesos biológicos en el cerebro, mientras que el otro se centró en la aplicación de redes neuronales a la inteligencia artificial. Este trabajo llevó a trabajar en las redes nerviosas y su relación con los autómatas finitos. [22]

4.2.2 Perceptrón

En el aprendizaje automático, el perceptrón es una red neuronal artificial formada por una sola neurona. Data de finales de los años cincuenta, fue la primera RNA y tiene una regla de aprendizaje que se describe a continuación.

Es un tipo de clasificador lineal, es decir, el algoritmo de clasificación hace sus predicciones basadas en una función de predicción lineal combinando un conjunto de pesos con el vector de características. El algoritmo permite el aprendizaje en línea, ya que procesa elementos en el conjunto de entrenamiento uno a la vez.

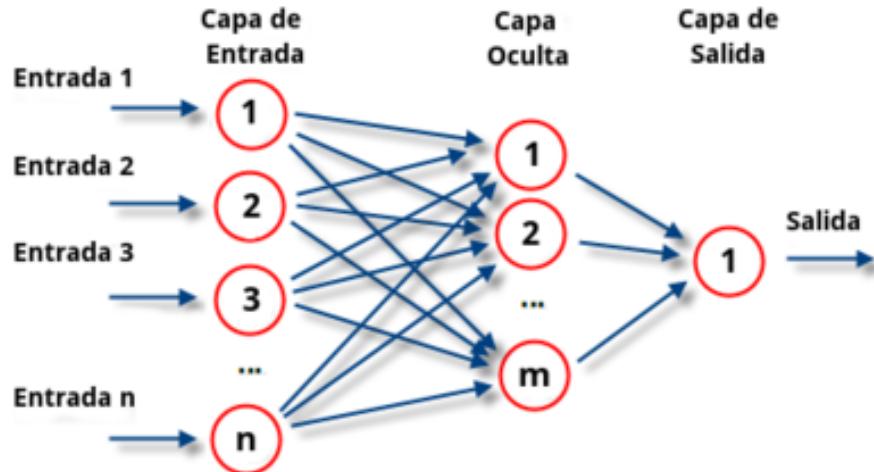


Figura 4.3: Arquitectura Feedforward

En la figura 4.2 se muestra un perceptrón que tiene tres entradas, x_0, x_1, x_2 hasta x_n . En general, podría tener más o menos entradas. Rosenblatt propuso una regla simple para calcular la salida. Introdujo pesos, w_0, w_1, w_2 hasta w_n (números reales) que expresan la importancia de las entradas respectivas a la salida. La salida de la neurona, 0 ó 1, está determinada por si la suma ponderada $\sum_j w_j x_j$ es menor o mayor que algún valor umbral. Al igual que los pesos, el umbral es un número real que es un parámetro de la neurona.

Ese es el modelo matemático básico. Una forma en que se puede pensar en el perceptrón es que es un dispositivo que toma decisiones al evaluar las pruebas.

4.2.3 Redes Feedforward

Una red neuronal feedforward es una red neuronal artificial en la que las conexiones entre las unidades no forman un ciclo. Como tal, es diferente de las redes neuronales recurrentes.

En esta red, la información se mueve en una sola dirección, hacia adelante, desde los nodos de entrada, a través de los nodos ocultos (si los hay) a la dirección de la red neuronal hacia la salida. No hay ciclos ni bucles en la red. La figura 4.3 muestra esta arquitectura.

Algoritmo Backpropagation

El algoritmo de propagación hacia atrás o Backpropagation [31] se utiliza en las RNA feed-forward en capas. Esto significa que las neuronas artificiales se organizan en capas, y envían sus señales hacia adelante, y luego los errores se propagan hacia atrás.

El patrón de entrada de la red se usa como estímulo, este se propaga desde la primera capa a través de todas las capas de la red, hasta generar una salida. La señal de salida se compara con la salida deseada y se calcula una señal de error para cada una de las salidas.

Las salidas identificadas como error se propagan hacia atrás, desde la salida, hacia todas las neuronas de la capa oculta. Este proceso se repite, capa por capa, hasta que todas las neuronas hayan recibido la señal de error. Este proceso es importante ya que, a medida que se entrena la red, las neuronas de las capas intermedias se organizan a sí mismas.

Puede haber una o más capas ocultas intermedias. El algoritmo de backpropagation utiliza el aprendizaje supervisado. Las RNA que implementan el algoritmo de backpropagation no tienen demasiadas capas, ya que el tiempo de formación de las capas crece exponencialmente.

4.3 Red neuronal Energía Coulombica Restringida (RCE)

La red neuronal RCE (Energía Coulombica Restringida) fue diseñada con el propósito de clasificar patrones de movimiento. Es capaz de resolver problemas de reconocimiento de patrones de clases lineales y no linealmente separables. En la literatura pueden encontrarse trabajos que la utilizan para la segmentación de color [50] [36].

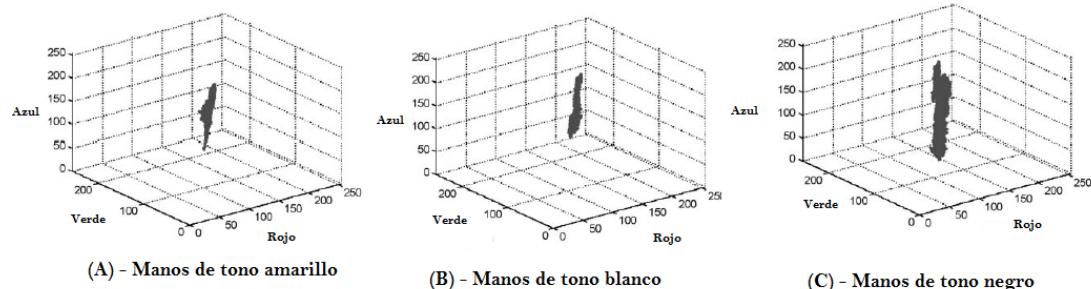
Energía Coulombica Restringida significa: columbica viene de culombio=Unidad de carga eléctrica, de símbolo C, que equivale a la cantidad de electricidad que transporta una corriente de intensidad de 1 ampere en 1 segundo.

Esta red es útil para la detección de colores de piel ya que en comparación con otros métodos, da mejores resultados. Este método nos permite detectar rápidamente sectores de la imagen donde podemos encontrar pixel de piel que pueden ser parte de la mano de una persona. El estudio sobre las distribuciones de color de la piel en RGB, HSI y L*A*B* [50] indica que los colores de la piel humana se agrupan en una pequeña región de un espacio de color y tienen más diferencia en intensidad que en color. En la figura

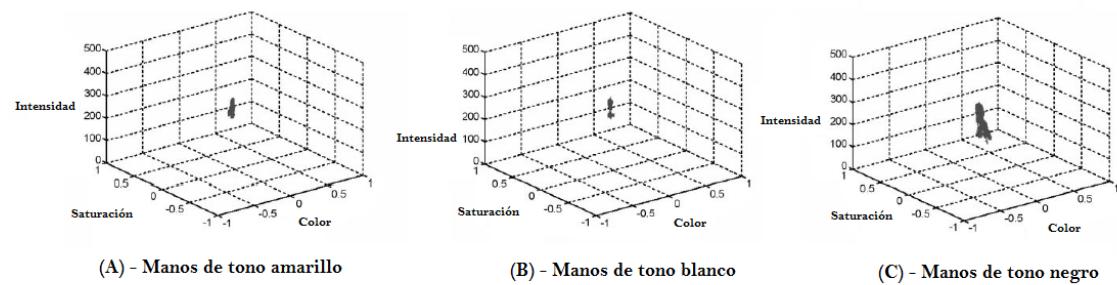
4.3. RED NEURONAL ENERGÍA COULOMBICA RESTRINGIDA (RCE)

4.4 se muestran las distribuciones de los pixeles de piel humana con distintos tonos en los sistemas de color RGB, HSI y L*a*b*.

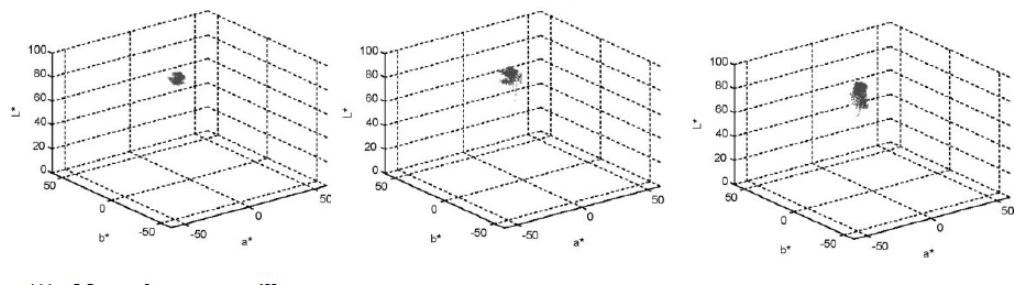
Durante el entrenamiento, la red neuronal RCE se caracteriza por agrupar con precisión las regiones de color de piel en el espacio de color, formando numerosas células donde se agrupan los conjuntos de colores de piel formando esferas de reconocimiento que contienen los diferentes colores e intensidades. Durante el entrenamiento de la red RCE se identifican todos los píxeles con tonos de piel en la imagen. Los resultados experimentales demuestran que este método de segmentación puede segmentar varias imágenes de mano eficientemente de todo tipo de fondos complejos en tiempo real.



Distribuciones de colores de piel en el espacio de color RGB



Distribución de colores de piel en el espacio de color HSI



Distribución de colores de piel en el espacio de color L*a*b*

Figura 4.4:

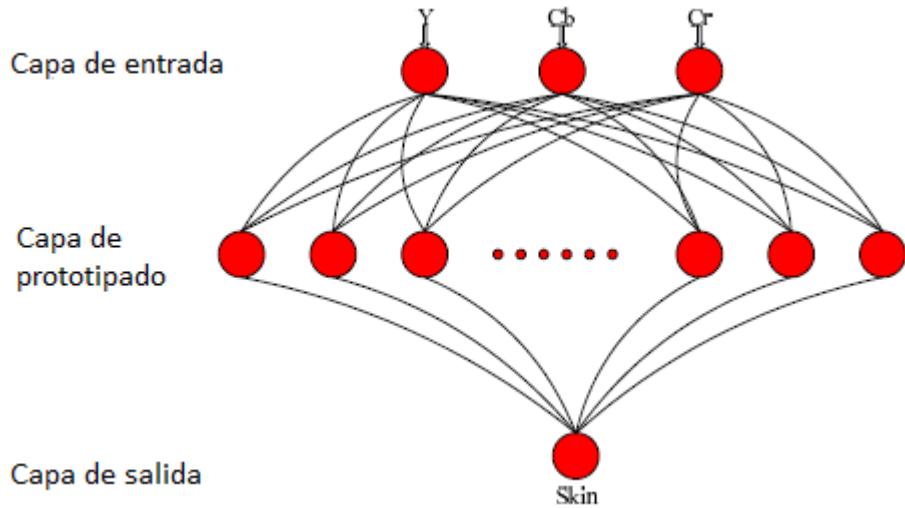


Figura 4.5: Arquitectura RCE

La arquitectura de la red RCE presenta tres capas de neuronas, la capa de entrada, la capa oculta, y la capa de salida. Las dos primeras están conectadas totalmente mientras que la oculta y la de salida tienen conexión parcial, como se muestra en la figura 4.5. La capa de entrada es la encargada de recibir los valores del espacio de color utilizado, como podría ser RCE, HSV, Y Cb Cr, etc. La capa intermedia o capa de prototipado contiene información de color de los pixeles conocidos como pixeles de piel que fueron incorporados en el aprendizaje. La ultima capa obtiene la clase correspondiente a la información de color del pixel entrante. Esto permite generar una clasificación de los pixeles ingresados como entrada.

Cada neurona de la capa oculta: se caracteriza por cinco elementos: clase C, peso vectorial ω , umbral celular λ , recuento de patrones κ y factor de suavizado θ . El vector de peso ω representa los pesos de las conexiones que la unen con cada neurona de entrada y por lo tanto tiene la misma dimensión que la señal de color. Un vector de peso define así un punto en el mismo espacio de color. El umbral λ describe una región esférica de influencia alrededor de la célula prototipo en el espacio de color.

En respuesta a una señal de color de entrada $X=(X_1, X_2, X_3)$, cada célula prototipo calcula una distancia entre la señal de color de entrada y el vector de prototipo de color almacenado en su $\omega = (\omega_1, \omega_2, \omega_3)$ como sigue:

$$(4.1) \quad (d_i) = \left[\sum_{j=01}^3 (\omega_{ij} - X_j)^2 \right]^{\frac{1}{2}}$$

donde ω_{ij} es el peso que conecta la i -ésima célula prototipo a la j -ésima célula de entrada, x_j el j -ésimo valor de color de la señal de entrada X . Una célula prototipo se activará para activar su clase de color asociada C si la distancia d del patrón al prototipo es menor que el umbral celular λ . Si d es mayor o igual que λ , el prototipo no responderá a la señal de entrada. Haciendo referencia a la salida de la i -ésima célula prototipo como π , esto significa:

$$(4.2) \quad P_i = \begin{cases} 1, & \text{si } d_i < \lambda_i \\ 0, & \text{si } d_i \geq \lambda_i \end{cases}$$

El recuento de patrones k almacena el número de veces que la célula prototipo se ha disparado correctamente en respuesta a las señales de entrada pertenecientes a su clase asociada. Es utilizado por la red para aproximar el valor de densidad de probabilidad local para una clase de color dada en el problema de la distribución de color no separable.

4.3.1 Entrenamiento de la red

Durante el algoritmo de entrenamiento se recorren los pixeles de entrada y se separan en neuronas. Esto se realiza a través del cálculo de la distancia entre los pixeles. Al finalizar la clasificación se evalúa la densidad de cada neurona, si la densidad de la misma no llega al mínimo esperado de pixeles, se descarta y se vuelven a reclasificar los mismos. Esto sucede hasta que dejen de aparecer nuevas neuronas por interacción. Se muestra el pseudo-código en el algoritmo 4.3.1.

$$(4.3) \quad d = \sqrt{(X_1 - X_2)^2 + (Y_1 - Y_2)^2}$$

$$(4.4) \quad D_{p^j} = \frac{3N^j}{4\pi R(j^3)}$$

El procedimiento de entrenamiento de RCE hace uso de distintos mecanismos: prototipo de compromiso celular e incremento del número de patrones de prototipo. Durante el entrenamiento, la red RCE clasifica cada uno de los patrones de entradas en la célula correspondiente, agrega la cantidad de células necesarias para cubrir las regiones de distribución para cada clase de color presentada en los datos de entrenamiento.

La Figura 4.6 muestra la región de distribución de los colores de piel construidos por células de prototipo y sus campos de influencia esférica en el espacio de color RGB.

Algoritmo 1 Algoritmo de entrenamiento de la red neuronal RCE

```

Setear el radio  $R_j = r$  (para todo j)
while sea la primer interacion o no se hayan creado nuevas neuronas en la interacion
anterior do
    for i = 1 a m(todos los elementos de entrenamiento) do
        if el pixel no esta etiquetado then
            for j = 1 a n(todas las neuronas) do
                calcular la distancia  $d(X_i, C_j)$  por la ecuación 4.3
                if  $d(X_i, C_j) \leq R_j$  then
                    incrementar  $N_j$  en 1, marcar  $X_i$  que pertenece a  $P^j$ 
                    salir del loop
                end if
            end for
            if  $d(X_i)$  no cayo en ninguna neurona existente then
                crear una nueva neurona  $P^{(n+1)}$  centrada en  $d(X_i)$  con radio r y  $d(N_j) = 1$ 
            end if
        end if
    end for
    for j = 1 a n(todas las neuronas) do
        calcular el valor de densidad de la neurona j con la ecuación 4.4
        if el valor de la densidad es mayor a  $\lambda$  then
            guardar la celda  $P^j$  y etiquetar todos elementos del entrenamiento como  $P^j$ 
        else
            descartar  $P^j$ 
        end if
    end for

end while

```

El estudio de la distribución del color de la piel ha indicado que los colores de la piel se agrupan en una pequeña región en un espacio de color, sin embargo, las regiones de distribución del color de la piel es complicada e irregular. Las técnicas más comunes de segmentación de color están basadas en histogramas, pero no son lo suficientemente efectivas para segmentar la imagen de una mano de un fondo complejo y dinámico, debido a la dificultad para seleccionar correctamente el umbral. La red RCE es capaz de segmentar imágenes de manos bajo condiciones de iluminación variable y fondos complejos, después de haber sido entrenada adecuadamente.

Una de las tareas más importantes es determinar cuales son los valores de los parámetros de la red, densidad y radio. El radio determina cual es el espacio que cubre/ocupa una neurona y que pixeles pertenecen a la misma. Si se establece un valor

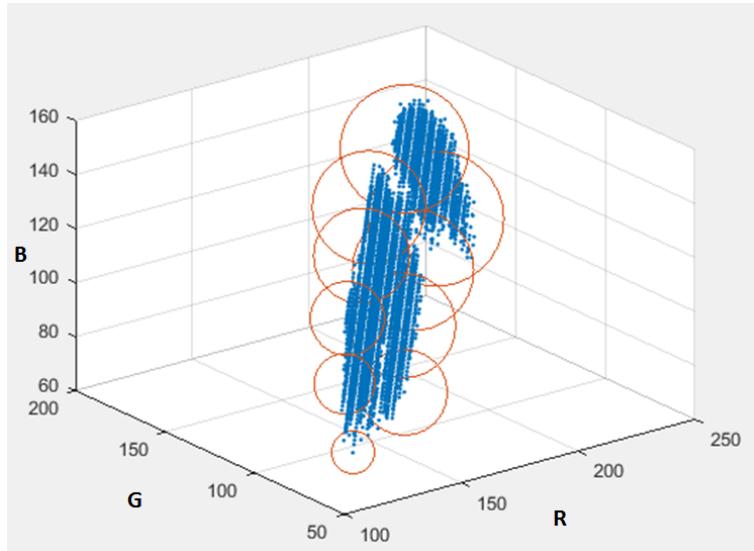


Figura 4.6: Muestra de resultados de la ejecución del algoritmo en 3D en el sistema de color RGB

de radio muy grande se obtiene menos cantidad de neuronas ya que se incrementa el espacio cubierto y además la capacidad de procesamiento es menor. Por otro lado, teniendo neuronas de radio muy grande se corre el riesgo de tener falsos positivos, es decir gran cantidad de pixeles que no pertenecen al color de piel que se está buscando pero por el valor de radio quedan comprendidos en el espacio de la neurona. Ahora si se tiene un radio chico se baja la probabilidad de tener falsos positivos pero la cantidad de neuronas aumenta considerablemente y ademas esto incrementa la velocidad de procesamiento. La densidad define cual es la cantidad mínima de patrones que se esperan tener en un neurona. En este caso, para una densidad grande, se necesita una gran cantidad de patrones para que la misma se mantenga en la red, pero esto asegura que esa neurona representa un gran numero de pixeles encontrados. En el caso de definir una densidad pequeña, permite que con una pequeña cantidad de pixeles la neurona sea aceptada. Ahora esto también permite tener muchas mas neuronas y aumenta la posibilidad de contener errores.

4.3.2 Clasificación

Luego del entrenamiento el proceso de clasificación de la red es sumamente sencillo. Dada una imagen I , se recorre la misma por completo, pixel a pixel realizando el mismo calculo de distancia que en el entrenamiento, de manera secuencial. Los pixeles se consideran

piel sólo si son clasificados en alguna de las neuronas entrenadas.

Cuando se tienen identificados los pixeles de color de piel en la imagen, se utiliza el método de *Template Matching* para verificar las formas de los objetos detectados y con esto concluir si es o no una mano.

Template Matching: es una técnica de procesamiento de imágenes digitales que busca encontrar partes de una imagen que coincidan con una imagen específica.

Este método permite encontrar similitudes entre las imágenes de entrada y las plantillas seleccionadas. Puede ser usado para determinar la correlación entre las imágenes de entrada y los patrones almacenados de todos los objetos que se desean detectar y determinar su ubicación. Para ello puede implementarse utilizando distintas técnicas como por ejemplo ¿distancia entre puntos?, correlación entre puntos o coeficiente, los cuales nos permiten encontrar características específicas.

La correlación es una herramienta importante tanto en el procesamiento de imágenes como en el reconocimiento de patrones y otros campos de investigación. La correlación entre señales es un principio fundamental para la detección de características como también para construir una técnica más sofisticada de reconocimiento. Para el caso de las manos se pueden tener tantos templates como formas de manos se desean detectar.

En esta técnica se utiliza un listado de plantillas pre-definidas que representan las formas de las manos aceptadas o reconocidas por la aplicación. Cada región del listado de candidatas se compara con cada plantilla pre-definida para determinar el grado de correlación que existe entre ambas. La forma de la mano reconocida es la que mayor valor de correlación tenga, siempre que se supere un umbral establecido como parámetro de configuración.

4.4 Limitaciones

Para construir un sistema de reconocimiento de gestos hechos con las manos robusto, se deben tener en cuenta diversos problemas. Uno de los principales es que la forma de la mano es difícil de caracterizar. A diferencia de la cara, por ejemplo, donde se pueden utilizar marcadores naturales como son los ojos y la boca para determinar si lo que se está procesando es efectivamente una cara, en caso de las manos, tanto abierta, cerrada, o con distintas posiciones de los dedos, se dificulta encontrar una características en común entre ellas, con lo cual, se convierte en un problema complejo, determinar si el objeto identificado, es una mano.

Por esto muchos autores utilizan el color de la piel para poder identificar si hay alguna parte del cuerpo de una persona en la imagen. Pero esto no es suficiente, ya que, se encuentran varios problemas, como se sabe, existen infinidad de tonalidades de piel entre las personas, y además el color de la piel en la mano, no es uniforme.

Al utilizar imágenes digitales de dos dimensiones nos encontramos con varias dificultades:

- Profundidad: Las técnicas de 3 dimensiones permiten que las soluciones no se vean influenciadas por la distancia en que se encuentra una persona realizando el gesto hacia el dispositivo, ya que cuentan con información adicional de la distancia donde se esta realizando el gesto. En el caso de las imágenes de dos dimensiones a mayor distancia respecto del dispositivo, el tamaño de la persona es menor, lo cual introduce mayor complejidad a la hora de comparar cual es el gesto y movimiento que se esta realizando.
- Iluminación: El color de los objetos dentro de una imagen se ven fuertemente afectados por la iluminación del ambiente al momento de la captura, esto tiene gran impacto a la hora de analizar cada uno de los pixeles de la imagen y determinar si en la misma se encuentra por ejemplo una mano.

4.5 Conclusión

En este capítulo se han describió que son las redes neuronales, como se desarrollan las redes neuronales artificiales. Además se detallo en detalle la arquitectura de la red neuronal RCE y como se utiliza la red para realizar el reconocimientos de patrones de piel en los pixeles de una imagen digital.

Ya que la red neuronal RCE nos permite obtener los colores de piel de las personas dentro de una imagen, se utilizo para desarrollar el algoritmo del prototipo desarrollado en esta tesina, en el capítulo siguiente se describe como fue implementado.

APLICACIÓN EN CONTROL DE TV

5.1 Introducción

Luego de haber analizado posibles alternativas que permiten segmentar y reconocer las manos de una persona en una imagen o vídeo, es posible utilizar esta información para generar un software con el cual un usuario pueda controlar un TV mediante gestos hechos con sus manos.

Este capítulo construye el aporte central de esta tesina y se describe en detalle como se integraron soluciones a las distintas partes del problema de reconocimiento de gestos hechos con las manos. El desarrollo de este proceso permite obtener como resultado un dispositivo capaz de capturar gestos y enviar señales infrarrojas a un TV con el objetivo de controlarlo.

Para ello es necesario tener en cuenta que este tipo de dispositivos, llamados embebidos se componen de dos partes fundamentales: el hardware y el software. Para la construcción de un prototipo capaz de detectar y realizar el seguimiento de las manos para ejecutar acciones del usuario, se combinaron soluciones de distintas áreas entre las que se destacan: redes neuronales, procesamiento de imágenes y electrónica.

En este capítulo se describen tanto el hardware como el software propuesto para una primera implementación.

Tanto el hardware como el software necesitan cumplir ciertos requerimientos que se detallarán a continuación para que el dispositivo pueda ser realmente útil. Es necesario detenerse en el detalle de la implementación de este prototipo para poder encontrar

una solución lo suficientemente robusta y escalable para ser capaz de adaptarse a otros dispositivos electrónicos.

En este capítulo se incluyen también las pruebas realizadas para validar el funcionamiento del software del prototipo. Estas abarcan tanto la verificación de la segmentación realizada por la red neuronal como la verificación de los resultados obtenidos por el dispositivo al realizar el procesamiento de los gestos.

5.2 Parte Uno - Hardware

Una parte fundamental en la construcción de este prototipo fue la elección de los componentes de hardware para obtener un dispositivo embebido que respondiera tanto a los requerimientos de software como a los de futuras ampliaciones de hardware. En este sentido, los requerimientos que debía cumplir el dispositivo a desarrollar son los siguientes: bajo costo económico, capacidad de cómputo para procesamiento de imágenes, capacidad de comunicación con otros dispositivos a bajo nivel, tamaño reducido, bajo consumo de energía, capacidad de funcionar con baterías y capacidad de actualización de hardware y software.

5.2.1 Análisis de posibles soluciones

Para llevar a cabo la construcción de dicho dispositivo es necesario analizar cuales son las alternativas disponibles en el mercado que permitan desarrollar un dispositivo con capacidad de computo en tiempo real para capturar imágenes, reconocer los gestos, realizar su seguimiento y enviar los comandos al TV.

Como se ha mencionado anteriormente, existen algunas soluciones desarrolladas por distintas empresas a este problema. En el caso específico del control de TV existe por ejemplo un producto de la empresa Samsung con su TV smart de control de interacción por movimiento (tv Samsung Smart Interaction Motion Control). La desventaja de estos dispositivos es que ya se encuentran integrados en los productos con lo cual no es posible extender el comportamiento o conocer de que manera realizan la segmentación y reconocimiento. Algunos teléfonos celulares también permiten realizar la captura de fotografías mediante el posicionamiento de la palma de la mano delante de la cámara en el momento que se desea tomar la fotografía. Estas soluciones están empaquetadas en productos y tienen como desventaja que no pueden ser utilizados para generar nueva funcionalidad.

5.2.2 Integración

Dado que el desarrollo del prototipo planteado en esta tesina puede definirse como un sistema embebido vale la pena describir algunas definiciones:

Sistema embebido

Un sistema embebido es un sistema informático con una función específica dentro de un dispositivo electrónico, es decir, que se ha incorporado el software en el hardware, para realizar una funcionalidad específica. Es decir, estos sistemas dispositivos con un hardware y software requerido para una tarea determinada. Los usos mas comunes de los sistemas embebidos son en los sistemas de tiempo real.

Está integración como parte de un dispositivo completo a menudo incluye hardware y otros dispositivos electrónicos necesarios para el software como pueden ser cámaras, sensores o componentes electrónicos. Los sistemas integrados controlan muchos dispositivos de uso común en la actualidad y son de gran utilidad para nuevas soluciones tecnológicas.

Control Remoto

En electrónica, un control remoto es un componente de un dispositivo electrónico utilizado para operarlo de forma inalámbrica desde la distancia. Por ejemplo, se puede usar un control remoto para operar dispositivos como un televisor, un reproductor de DVD u otro electrodoméstico, desde una distancia corta. Un control remoto es principalmente un elemento que permite al usuario operar dispositivos que están fuera de su alcance, o no tiene acceso directo de los controles. En algunos casos, los controles remotos permiten que una persona opere un dispositivo que de otro modo no podría alcanzar, como cuando un abre-puertas de garaje se activa desde el exterior.

Los primeros controles remotos de televisión (1956-1977) usaban tonos ultrasónicos. Los controles remotos actuales son comúnmente dispositivos de infrarrojos que envían pulsos codificados digitalmente de señales infrarrojas para controlar funciones tales como potencia, volumen, ajustes, ajuste temperatura, velocidad del ventilador u otras características. Los controles remotos para estos dispositivos suelen ser pequeños objetos de mano, inalámbricos, con una serie de botones para ajustar diversas configuraciones, como el canal de televisión, el número de pista y el volumen. Para muchos dispositivos, el control remoto contiene todos los controles de función, mientras que el dispositivo controlado sólo tiene un puñado de controles primarios esenciales. El código de control

remoto, y por lo tanto el dispositivo de control remoto requerido, es generalmente específico para una línea de producto, pero hay controles remotos universales, que emulan el control remoto hecho para la mayoría de los dispositivos de marca.

Un sistema de control remoto actual consiste en dos partes importantes, un emisor y un receptor de señales infrarrojas. El emisor debe ser capaz de transmitir un código de bits el cual debe ser conocido por el receptor. Cada código de bits trasmisido comprende una señal de datos que representa un estado lógico seleccionado seguido inmediatamente por una segunda señal de datos que representa el estado lógico complementario.

Señales Infrarrojas

La comunicación IR o infrarroja es una tecnología de comunicación inalámbrica común, económica y fácil de usar. La luz IR es muy similar a la luz visible, excepto que tiene una longitud de onda ligeramente más larga. Esto significa que IR no es detectable para el ojo humano, lo que lo hace útil para la comunicación inalámbrica. En el anexo A se describe en detalle como funcionan las señales infrarrojas.

Las aplicaciones de espacio de usuario le permiten controlar su computadora con su mando a distancia. Puede enviar eventos X11 a aplicaciones, iniciar programas y mucho más con sólo pulsar un botón. Las aplicaciones posibles son obvias: Mouse infrarrojo, control remoto para su tarjeta sintonizadora de TV o CD-ROM, apagado por control remoto, programe su videograbadora y / o sintonizador de satélite con su computadora, etc. Utilizando lirc en Raspberry Pie es bastante popular estos días.

5.2.3 Construcción del prototipo

Teniendo en cuenta los requerimientos de hardware antes mencionados se decidió utilizar una Raspberry Pi 3 como corazón del dispositivo. Cuenta con 1Gb de memoria RAM y un procesador ARM Cortex-A53 de 1.2GHz con 4 núcleos lo que ofrece un nivel de cómputo aceptable para procesamiento de imágenes moderado. También tiene un tamaño reducido, capacidad de comunicación con otros dispositivos a través de su GPIO y de sus 4 puertos USB. Cuenta con comunicación inalámbrica vía WiFi y Bluetooth. Requiere una alimentación de 5V y 1.5A para nuestro tipo de aplicación, lo que hace posible el uso de baterías, como por ejemplo un powerbank. Todo esto a un costo razonable.

Para capturar los gestos de las manos se utilizó una cámara web convencional con una tasa de adquisición de unos 20 cuadros por segundo a una resolución de 1280 x 1024

píxeles. Esta cámara permite capturar imágenes de dos dimensiones en el sistema de color RGB.

Para controlar las funciones del TV se utilizó un módulo emisor de luz infrarroja que permite codificar los comandos que lo controlan. También se incluyó un módulo receptor infrarrojo que permite incorporar al sistema los códigos de aquellos controles remotos que tengan códigos desconocidos o no muy populares.

También se utilizó un módulo para sensar la luz ambiente con el objetivo de realizar ajustes en la luminosidad de la imagen capturada con la cámara web. Este sensor permite obtener cual es la intensidad de la luz en el momento que se están capturando las imágenes, en lux (unidad de medida del nivel de iluminación).

En la figura 5.1 se pueden ver las imágenes de estos dispositivos.

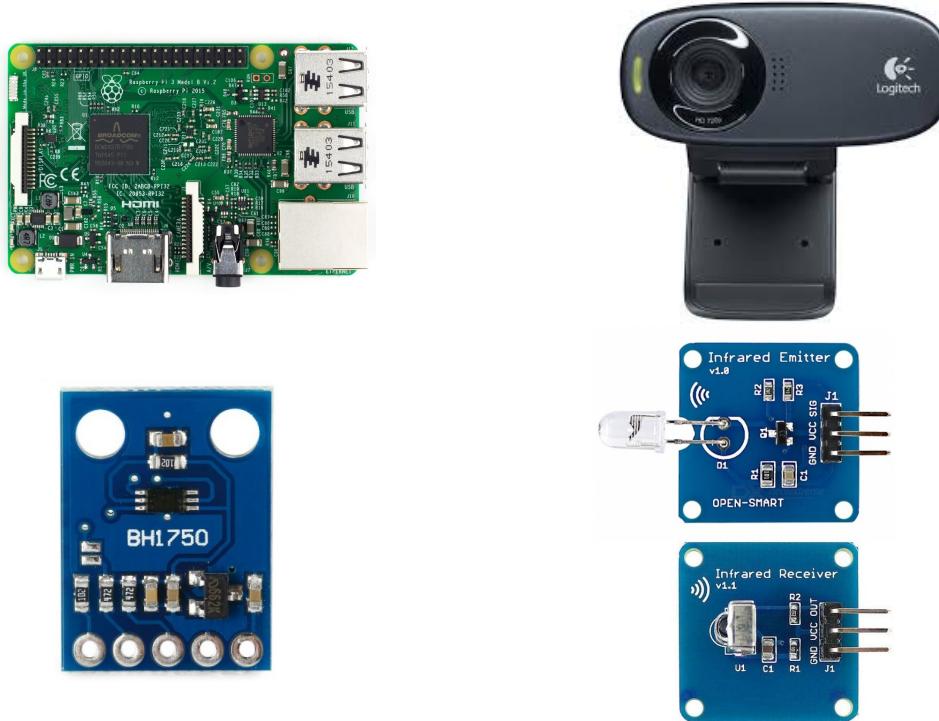


Figura 5.1: Módulos que componen el hardware. (a) Raspberry PI 3. (b) Cámara web. (c) Sensor de luminosidad. (d) Emisor y receptor infrarrojos.

Podemos mencionar algunas ventajas de utilizar una raspberry pi para desarrollar este prototipo, este dispositivo tiene muy bajo costo, cuenta con un tamaño reducido, permite expansión (Wifi, Bluetooth, Entrada/Salida de bajo nivel, HDMI, Audio), código libre, tiene capacidad de cómputo para procesamiento de imágenes y ademas cuenta con una gran comunidad la cual nos permite tener soporte de manera activa.

5.3 Parte dos - Software

Como se ha mencionado este prototipo se compone de dos partes fundamentales, hardware y el software. En el caso del software tenemos una combinación de herramientas utilizadas.

El primer paso de este proyecto consistió en la búsqueda de una alternativa que permita la segmentación de las manos de una persona dentro de una imagen digital. Como se estudiaron y analizaron varios problemas, ya mencionados, que comúnmente se encuentran en el procesamiento de imágenes (diferencias en colores, problemas de iluminación, distancia a la cámara, etc). Como resultado de esta búsqueda se decidió utilizar la red neuronal RCE descripta en el capítulo 4, debido a que en las pruebas realizadas sobre distintos sistemas de representación color se obtuvieron resultados satisfactorios para segmentar diferentes tipos de piel.

Al elegir esta técnica fue necesario construir una base de datos de colores de piel para determinar la efectividad de la misma. Se tomaron imágenes con una cámara digital convencional, en condiciones de iluminación y distancia idénticas y se capturaron al menos 16 posiciones distintas de las manos, de 8 individuos con tonos de piel diferentes.



Figura 5.2: Configuraciones

Estas imágenes se sometieron a pruebas con redes neuronales RCE entrenadas en distintos sistemas de representación de color, rgb, hsl, hs y Cie-Lab. El objetivo de las pruebas fue comparar las distintas redes en cada sistema para determinar en cual de ellos se obtienen mejores resultados para este problema específico. A fines prácticos también se realizaron pruebas similares con la base de datos de manos "MOHI" y con una base de datos de imágenes que no contienen piel.

Esta base de datos se utilizó para entrenar una red neuronal de arquitectura dinámica, con la cual se implementó un reconocedor de colores de piel basado en redes neuronales, usando como entrenamiento inicial esa BDD.

Para este reconocedor de piel inicial se permitió realizar ajustes a través de un proceso de calibración inicial para compensar problemas de iluminación.

Este primer reconocedor de colores de piel fue implementado en el lenguaje Matlab. Se realizó un prototipo de la red neuronal RCE donde eso proceso la BDD inicial, mejorando el algoritmo para poder reducir los tiempos de ejecución, ya que, al realizar el procesamiento de cada pixel de las imágenes capturadas y teniendo en cuenta que una cámara promedio captura alrededor de 25 frames por segundo el procesamiento matricial de las imágenes era realmente costoso.

Este prototipo luego fue migrado a una implementación en el lenguaje Python utilizando Open CV para el procesamiento de imágenes y bibliotecas como Numpy y Scipy para el cálculo de matrices.

5.4 Software para Reconocimiento de Gestos

Teniendo en cuenta que esta es la primera versión del dispositivo, se decidió limitar su funcionamiento a una aplicación relativamente simple que no demandara una gran cantidad de cómputo. De esta manera se construyó un prototipo funcional tanto de hardware como de software que permite reemplazar el control remoto de un televisor convencional por un controlador que reciba las instrucciones a través de gestos realizados con las manos. En este sentido se limita tanto el modelo de representación como la cantidad de gestos de la mano, aunque en futuras versiones se tiene planificado ir incorporando un modelo más complejo como el planteado aquí [30]. Respecto del software utilizado para el desarrollo de este prototipo, se utilizó la versión Jessie de Raspbian, que es el sistema operativo con soporte oficial de Raspberry. El sistema fue implementado en Python por su facilidad para prototipado rápido y por la gran disponibilidad de bibliotecas que ofrecen tanto algoritmos tradicionales como de vanguardia. Para las partes que requirieron procesamiento de imágenes se utilizó la biblioteca OpenCV que tiene una sólida madurez y sus algoritmos están altamente optimizados, incluso para aprovechar las características de la GPU. Tanto para el entrenamiento y funcionamiento de la red neuronal RCE como para el algoritmo de seguimiento de la mano se utilizó una implementación propia optimizada para NumPy.

A continuación se describen en detalle los pasos de cada una de las etapas en las que se divide todo el proceso de reconocimiento del gesto para controlar un TV. En el esquema de la figura 5.3 se puede observar todo el proceso que realiza el software del dispositivo.

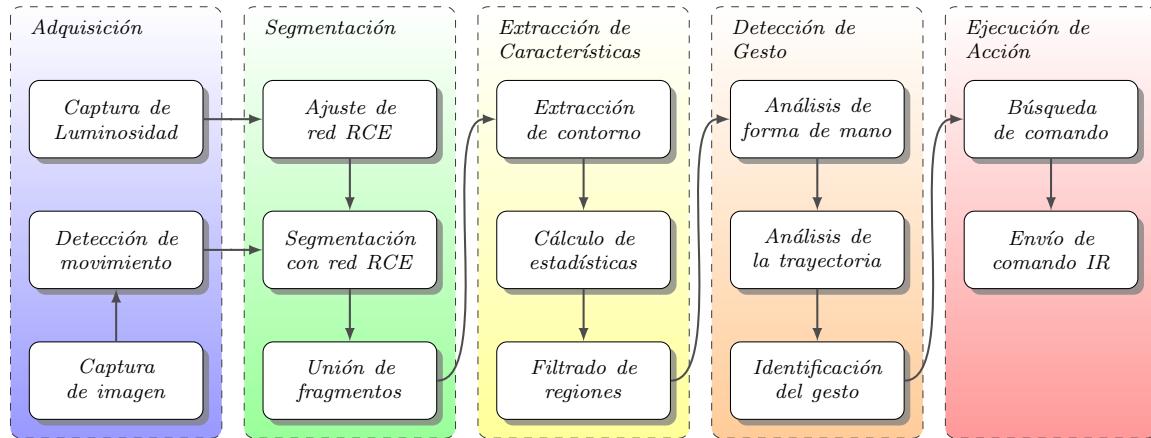


Figura 5.3: Proceso de reconocimiento del gesto y ejecución de comandos.

En el algoritmo 5.4 se describe en pseudo-código la implementación del prototipo desarrollado.

Algoritmo 2 Algoritmo de reconocimiento de gestos

```

iniciar cámara
obtener red entrenada
calibrar luz inicial
while cámara prendida do
    Obtener un frame
    Procesar el frame con RCE
    Calcular estadísticas
    Filtrar de regiones
    if existe una mano en la mascara then
        if Comparar si continua el movimiento o empezó then
            enviar la señal infrarroja
        end if
    end if
end while

```

5.4.1 Adquisición

La etapa de adquisición se divide en dos partes. La primera consiste en la captura, mediante una cámara web convencional, de una imagen RGB con una resolución de 1280 x 1024 píxeles a una tasa de adquisición que puede variar de 10 a 20 cuadros por segundo. Una vez obtenida la imagen correspondiente al cuadro actual, se la compara

con la del cuadro adquirido anteriormente para determinar si se produjo movimiento y así evitar el procesamiento innecesario.

La segunda parte consiste en la adquisición de la cantidad de luz ambiente a través de un transductor que mide esta variable en lux ($lumen/m^2$). Este valor obtenido permite ajustar la luminosidad de la imagen en la etapa de segmentación y de esta manera corregir las variaciones en la iluminación que afectan a los colores.

5.4.2 Segmentación

Para realizar la segmentación de la piel humana se ingresa cada pixel de la imagen a la red RCE para determinar si éste se corresponde o no al color de la piel. Como resultado de esta operación se obtiene una máscara preliminar a la que luego se le aplica una operación morfológica de cierre para unir áreas que pudieran haber quedado desconectadas.

Es importante destacar que antes de realizar la segmentación, con el objetivo de subsanar el problema de inestabilidad que provocan las variaciones de iluminación, se aplica un ajuste a la red RCE según la luz ambiente captada por el sensor de luminosidad. Esta corrección se aplica a los valores RGB de las neuronas de la capa intermedia para neutralizar los cambios de intensidad de luz que están presentes en la imagen.

En la figura 5.4 se muestra la máscara obtenida después de haber procesado la imagen capturada con la red neuronal RCE.



Figura 5.4: Imagen procesada con RCE

5.4.3 Extracción de características

En esta etapa, luego de obtenidas las áreas de la imagen donde se localiza la piel, se determina cuales de todas esas porciones pueden corresponderse con la mano de una persona.

El proceso inicia con la máscara que representa a todos los píxeles reconocidos como piel en la etapa anterior. A ésta máscara se le aplica un algoritmo de extracción de contornos para obtener un listado de estos. Cada contorno agrupa píxeles interconectados en la máscara que forman una región donde potencialmente podría haber una mano.

Una vez obtenido el listado, por cada contorno se calcula el área, el perímetro y los ejes principales que son propiedades que dan una pauta de las características geométricas generales de la región.

Finalmente, para filtrar el listado, se analiza cada contorno o región comparando sus propiedades con las propiedades de una región que contiene una mano, descartando aquellos que difieren mucho de lo esperado. Como resultado de esta comparación se obtiene un listado de contornos candidatos que representan regiones potenciales donde puede encontrarse una mano.

En la figura 5.5 se muestra la máscara que fue recortada después de ser procesada la figura 5.4.



Figura 5.5: Mano Recortada

5.4.4 Detección del gesto

Para determinar si las regiones candidatas obtenidas en la etapa anterior se corresponden o no con una mano se aplica la técnica de *Template Matching*.

Una vez que se obtiene una región candidata que coincide con alguna de las formas pre-definidas de mano, se inicia el proceso de seguimiento tomando como referencia la posición de la región como posición de inicio. Luego se analiza la secuencia posterior de imágenes localizando la mano para determinar la evolución de su posición. Cuando

la distancia entre la posición inicial y la posición actual de la mano supera un valor determinado por un parámetro de configuración se establece la dirección del movimiento y se procede a determinar si el gesto coincide con los definidos en la aplicación.

En la figura 5.4.4 se muestran dos imágenes en las cuales se puede observar del lado izquierdo el frame capturado por la cámara y del lado derecho el resultado de haber procesado el frame con la red neuronal. Ademas de puede ver en el frame marcado con un recuadro rojo cual es el recorte de la imagen donde el algoritmo detecto una mano, y en ello se puede observar remarcado en verde el contorno de la mano detectada.

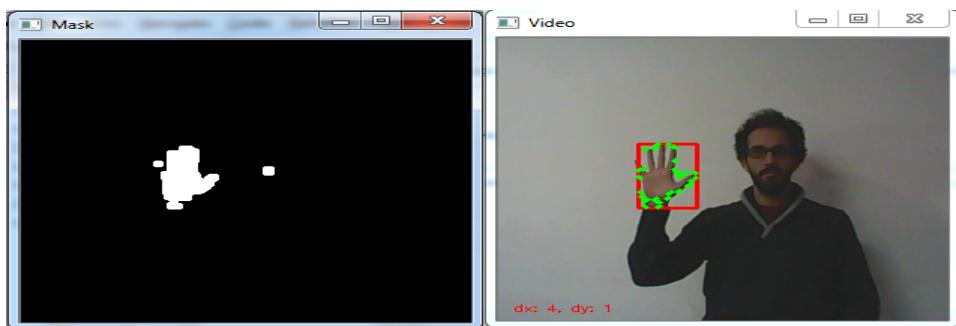


Figura 5.6: Ejemplo mano abierta

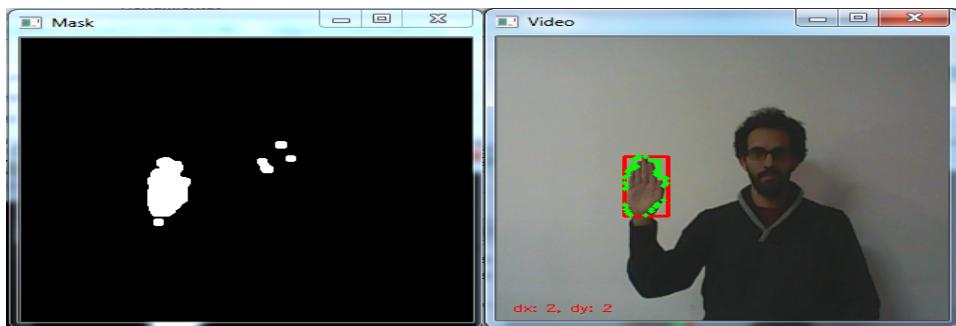


Figura 5.7: Ejemplo mano cerrada

5.4.5 Ejecución de Acción

Identificado el gesto de la mano, se determina el comando infrarrojo que éste tiene asociado para enviarlo al TV. Para ello, se utiliza un pequeño módulo de hardware que genera una señal infrarroja que es recibida por el TV, quien la decodifica e interpreta para ejecutar la función asociada a dicha señal.

Para codificar los comandos infrarrojos se utiliza la biblioteca LIRC (Linux Infrared Remote Control). Esta biblioteca permite decodificar y reproducir una secuencia de

pulsos infrarrojos de la misma manera que lo hace un control remoto convencional. Cuenta con una gran cantidad de códigos predefinidos de controles remotos e incorpora un servicio con la capacidad de copiar, desde un control remoto, aquellos códigos que no están predefinidos.

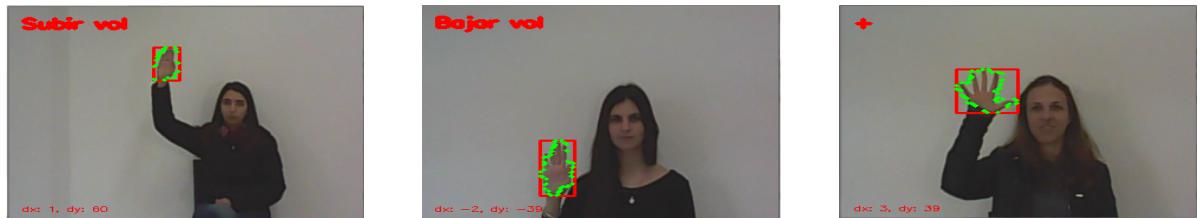


Figura 5.8: (a) Subir el volumen. (b) Bajar el volumen. (c) Siguiente canal

En la figura 5.4.5 se pueden observar tres ejemplos de las pruebas realizadas. En la figura a se puede ver que al realizar el gesto de subir la mano cerrada hacia arriba se envía el comando al TV de subir el volumen. En la figura b se realiza el gesto de bajar la mano cerrada y se envía el comando al TV de bajar el volumen. Y por ultimo en la tercera figura c se realiza el gesto de subir la mano abierta y se envía el comando al TV de aumentar el numero de canal.

5.5 Pruebas/Resultados

Luego de armado el prototipo se realizaron diferentes pruebas para evaluar la eficacia del mismo, de los cuales es necesario tener en cuenta los siguientes aspectos:

Tipos de piel: Es sabido que los tonos y texturas de piel de las personas son muy diferentes. Utilizar una técnica de detección basada en umbrales fijos es bastante complejo si tenemos en cuenta esta información. Al utilizar la red neuronal RCE podemos contemplar mejor este problema pero es necesario que la misma esté lo suficientemente entrenada para no dejar colores de piel sin detectar. Existen infinitos colores de piel en las personas, por lo que se vuelve bastante complejo determinar cuantos colores son necesarios para tener una red completa, que pueda identificar todos los tonos de piel. Este problema puede ser resuelto entrenando la red neuronal con una amplia gama de colores de piel posibles, lo cual permitirá tener mayor efectividad.

Problemas de iluminación: No solo los diferentes colores de piel son un problema a la hora de detectar objetos en una imagen. La iluminación también juega un papel importante en la misma. Los distintos tipos e intensidades de luz afectan a los tonos que puede llegar a tomar la piel en cada una de las imágenes. El procesamiento de

estas imágenes se ve afectado por los cambios de iluminación. Existen variaciones en la luminosidad de la imagen si la misma es tomada al sol o con iluminación artificial, estas también son variables ya que la luz puede ser blanca o amarilla, esto hace que cambien los valores de los pixeles de la imagen por lo que esta situación complica el procedimiento de detección.

Teniendo en cuenta estas consideraciones se realizaron las pruebas tanto para el procesamiento de imágenes con la red neuronal RCE, como también para el prototipo terminado, las cuales se describen a continuación.

5.5.1 Sistemas de Color y Red Neuronal RCE

Para la segmentación de manos se utilizó una red neuronal RCE [5, 28, 45] que determina cuando un pixel de la imagen se corresponde con el color de la piel. En la revisión de la literatura sobre la segmentación de piel basada en el color del pixel [5, 14, 41, 45] se encuentra que distintos algoritmos aplicados a imágenes en diferentes sistemas de representación del color obtienen resultados aceptables. Por este motivo, se decidió realizar una serie de pruebas en los sistemas de representación RGB, HSL, HSV, YCbCr, Cie-LAB para determinar cual es el más conveniente. Para realizar las pruebas en los distintos sistemas se utilizó la base de datos MOHI [11] que contiene muestras de manos de 250 personas y una base de datos construida ad hoc con imágenes que no contienen piel.

En la figura 5.9 se muestran los resultados obtenidos en las distintas pruebas realizadas. Para cada sistema de color se muestran dos barras que expresan el promedio de píxeles clasificados como piel cuando la imagen contiene piel y cuando no contiene piel. En general se puede observar que no hay grandes diferencias entre los distintos sistemas de color. También se puede observar que una mejora en la detección de los pixeles de la BDD con piel incrementa la cantidad de píxeles detectados en la BBDD sin piel y viceversa. Esto concluye que al incrementar la cantidad de pixeles detectados en un sistema de color también aumenta la tasa de errores detectados. En consecuencia, se decidió optar por utilizar el sistema RGB para la segmentación con la red RCE para evitar el costo del cómputo de la transformación a otro sistema debido a que la mejora no es significativa y a que el procesamiento se realiza en tiempo real, lo cual podría afectar en performance.

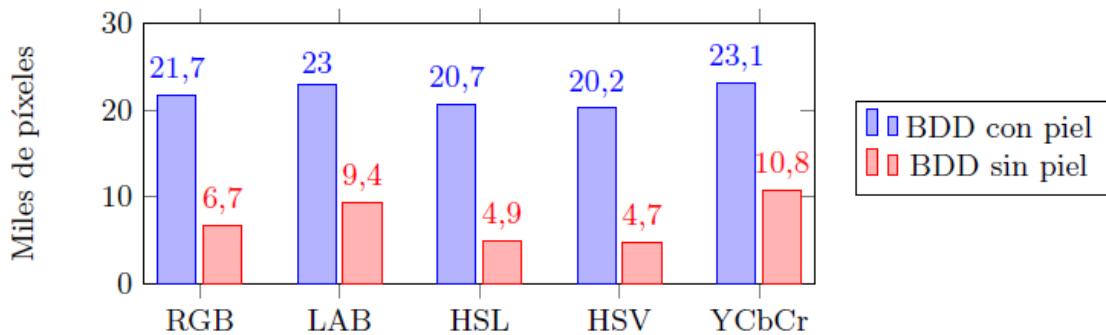


Figura 5.9: Red neuronal RCE en diferentes sistemas de representación del color.

5.5.2 Reconocimiento de Gestos

En principio este trabajo comenzó buscando técnicas para reconocer las manos utilizando solamente una cámara convencional. Se entrenó una red neuronal RCE para clasificar los píxeles de una imagen con el fin de discriminar si pertenece o no al color de piel humana. Junto con ello se procesaron las formas obtenidas para determinar si las mismas eran o no formas de manos de una persona.

Con el prototipo terminado se realizaron pruebas para medir su precisión en un ambiente controlado. Para esto se utilizaron 11 sujetos en dos condiciones diferentes de iluminación. Una condición de iluminación es baja (30 lux) y la otra condición de iluminación es media (al menos unos 600 lux). De cada sujeto se tomaron 3 formas de la mano (figura 5.5.2): mano abierta con dedos separados, mano abierta con dedos juntos y mano cerrada con índice y pulgar separados. Con cada forma se realizaron movimientos en dos direcciones opuestas (arriba y abajo).



Figura 5.10: Formas de manos. (a) Mano abierta con dedos separados, (b) Mano abierta con dedos juntos. (c) Mano cerrada con índice y pulgar separados.

En la tabla de la figura 5.11 se pueden observar los resultados obtenidos. Se definieron realizar gestos básicos para este primer prototipo generado, como por ejemplo mostrar la palma de la mano con los dedos abiertos hacia arriba, hacia abajo, también de la

misma manera pero teniendo los dedos juntos. Esto permite que sean gestos sencillos y poder medir si el objetivo del prototipo puede cumplirse. Cuando las formas de las manos empiezan a complejizarse o los movimientos se combinan empiezan a aparecer nuevas dificultades, como por ejemplo una de ellas puede ser las sombras. Como se menciono la primer segmentación de las manos se hace mediante una red neuronal RCE que permite obtener los pixeles que pertenecen al color de piel humana de una imagen, pero cuando los dedos de las manos se doblan aparecen sombras entre los mismos y eso hace que se puedan perder esos pixeles.

Gesto	Distancia 1,80 m		Distancia 3 m	
	Iluminación 1	Iluminación 2	Iluminación 1	Iluminación 2
Mano Cerrada- Arriba	100%	100%	100%	100%
Mano Cerrada - Abajo	100%	100%	100%	100%
Mano Abierta - Arriba	86%	45%	66%	33%
Mano abierta - abajo	71%	45%	100%	88%
Forma - Arriba	14%	0%	0%	0%
Forma - Abajo	14%	0%	0%	0%

Figura 5.11: Resultados en porcentaje de la detección de los gestos.

Para las pruebas con las formas de las manos se puede observar que para los casos de mano abierta se pueden detectar bien en condiciones de iluminación aceptable. Para los casos de mano abierta con dedos juntos se puede observar que falla cuando la iluminación es baja. Para los casos de la mano con índice y pulgar se observa que la detección falla de manera importante en las 2 condiciones de iluminación. La principal causa de este problema es la sombra que genera la flexión de los dedos. Esta sombra hace que la segmentación falle al no poder reconocerla como piel y se produzcan separaciones importantes que hacen que la región de la mano se extraiga parcialmente. Luego no se encuentran coincidencias o bien porque la región extraída se descarta porque no cumple con las propiedades geométricas esperadas o porque la técnica de *Template Matching* falla porque no es robusta para encontrar coincidencias parciales.

Luego de realizar el primer conjunto de pruebas surgieron algunas mejoras a realizar en el algoritmo para mejorar los resultados. Ademas se incorporaron las evaluaciones de las mismas pruebas con los movimientos de las manos hacia la derecha e izquierda. En la figura 5.12 se muestran los resultados d dichas pruebas.

Se puede observar que la evaluación de los gestos a la derecha fue satisfactoria. En el caso de la izquierda era esperable que el algoritmo falle ya que al moverse la

Gesto	Distancia 1,80 m		Distancia 3 m	
	Iluminación 1	Iluminación 2	Iluminación 1	Iluminación 2
Mano Abierta - Arriba	90%	80%	100%	100%
Mano abierta - abajo	90%	100%	100%	100%
Mano abierta - Derecha	100%	100%	85%	100%
Mano abierta - Izquierda	55%		45%	
Mano Cerrada- Arriba	100%	100%	100%	100%
Mano Cerrada - Abajo	100%	100%	100%	100%
Mano Cerrada - Derecha	90%	85%	55%	100%
Mano Cerrada - Izquierda	55%		30%	
Forma - Arriba	20%	75%	30%	70%
Forma - Abajo	40%	60%	20%	70%
Forma - Derecha	30%	45%	30%	35%
Forma - Izquierda	20%		0%	

Figura 5.12: Resultados en porcentaje de la detección de los gestos. Prueba 2

mano sobre la cara se detectan también los pixeles de color de piel, y la forma obtenida no corresponde a la forma esperada. En los casos en que los sujetos que realizan los movimientos no son realizados sobre la cara, los resultados fueron satisfactorios también. Esto sugiere algunos cambios que podrían salvar esta dificultad pero serán incorporados en una segunda etapa.

En las figuras 5.5.2 se muestran algunas imágenes de los gestos realizados en los vídeos procesados con los 4 movimientos, arriba, abajo, derecha e izquierda donde se puede observar en el margen superior izquierdo que el gesto fue reconocido satisfactoriamente.

5.6 Conclusión

En este capítulo se detalló cuales son los elementos que conforman el prototipo del dispositivo, como así también el software instalado en el para poder capturar los gestos realizados con las manos para convertirlos en señales interpretadas por un TV. Se describió en detalle como se realiza cada paso para realizar la segmentación, caracterización y el tracking, lo cual permite obtener la información de los gestos realizados y finalmente poder ejecutar acciones para enviarlas al TV.

También se describen las pruebas realizadas durante el desarrollo del prototipo, tanto de la red neuronal elegida como para el algoritmo implementado para el reconocimiento de gestos.

Las pruebas realizadas para la red neuronal permitieron definir que la conversión de la imagen obtenida en RGB a otro sistema de color no era necesaria ya que la mejora de



Figura 5.13: Gestos detectados por el algoritmo. (a) Abajo mano abierta. (b) Arriba mano cerrada. (c) Derecha mano abierta. (d) Izquierda mano cerrada.

la tasa de aciertos no era significativa.

Para el reconocimiento de gestos se describieron dos pruebas realizadas, una inicial la cual permitió realizar mejoras sobre el algoritmo y una segunda prueba la cual incorpore dichos cambios y algunos gestos las cuales dieron resultados satisfactorios para esta primer versión del prototipo.

Con los resultados obtenidos de estas pruebas se puede concluir que este prototipo inicial permite enviar las señales al TV dependiendo de los gestos realizados, es decir, el objetivo planteado fue completado.

En el próximo capítulo se expondrán las conclusiones del trabajo y las líneas de trabajo futuro.

CONCLUSIONES

En este trabajo se ha realizado una investigación sobre las posibilidades que existen para hacer reconocimiento de gestos hechos con las manos, tanto de productos cerrados como trabajos realizados por distintos autores en la actualidad. Se investigo en detalle cada etapa que compone la segmentación, captura, caracterización y reconocimiento del gesto.

Luego se presento el diseño de una primera versión de un prototipo de un sistema reconocedor de gestos realizados con las manos para controlar las funciones de un TV. Esta primera versión fue evaluada con una serie de sujetos de prueba y se obtuvieron resultados para poder evaluar las mejoras necesarias para la próxima versión del prototipo.

Resultados esperados

En esta tesina se esperan obtener los siguientes resultados:

- Construir una base de datos de colores de piel para entrenar una red neuronal de arquitectura dinámica, mediante una cámara convencional.
- Generar un BBDD de configuraciones de manos para medir el desempeño del método propuesto.
- Desarrollar un prototipo del controlador de TV utilizando un sistema embebido.

- Estudiar de las técnicas basadas en redes neuronales que pueden utilizarse para la segmentación de manos en un vídeo.

Entre los resultados que se obtuvieron se encuentran:

- Se obtuvo una base de datos de colores de piel para realizar el entrenamiento necesario con una gran cantidad de tonos diferentes.
- Se pudo obtener la segmentación de piel aceptable utilizando una red neuronal simple como es el caso de RCE, que ademas funciona independientemente del sistema de color con el que la imagen alla sido obtenida.
- Se pudo observar que la variación de la luz del ambiente todavía tiene impato sobre el prototipo. Este es un aspecto aún no resulto pero en este sentido se encontró una solución parcial a esta dificultad utilizando lecturas de un sensor de luz para adaptar la red neuronal RCE a las condiciones de iluminación del ambiente. Cuando hay ausencia de luz ambiente o esta es notablemente baja, la corrección por intensidad de la luz no resulta suficiente para una cámara web convencional. En este caso es conveniente realizar una adaptación de la cámara para que funcione con iluminación infrarroja.
- Se pudo utilizar harware con bajo costo y capaz de realizar el procesamiento de las imágenes capturadas por la cámara web y capaz de realizar el procesamiento necesario en ellas.

Este trabajo aporta una primera versión de un dispositivo de reconocimiento de gestos hechos con las manos que permite ejecutar un conjunto acotado de gestos capases de ser enviados a un TV. Este prototipo nos permite tener un producto base con el cual obtener las mejoras necesarias para poder ser ampliado y utilizado en otro tipo de electrodomesticos utilizados habitualmente.

6.1 Líneas de trabajo futuras

Debido a que este es un trabajo inicial, hay algunas líneas de trabajo que se están desarrollando y otras líneas que quedan por explorar y desarrollar. Actualmente se está trabajando para mejorar la segmentación de las áreas de la piel. Muchas veces debido a la combinación de iluminación con la posición de las manos, se producen sombras en las uniones de las falanges que provocan que la segmentación deje los dedos separados de las

6.1. LÍNEAS DE TRABAJO FUTURAS

manos. Una solución que se está explorando es la incorporación al entrenamiento de la red neuronal RCE ejemplos de piel con sombra, lo que permitiría mejorar la segmentación en la unión de las falanges. Los resultados obtenidos en algunas pruebas preliminares que se realizaron al momento de la escritura de este documento son prometedores en este aspecto.

En lo que se refiere al hardware, aún resta explorar las maneras de controlar las funciones distintos dispositivos electrónicos mediante red cableada, red inalámbrica, bluetooth e infrarrojo para aprovechar las capacidades del dispositivo y expandirlas. Finalmente quedan por analizar algunas alternativas de reemplazo de la Raspberry Pi 3 para reducir el costo. Dispositivos como Orange Pi, en sus versiones Zero, One y Lite podrían reducir el costo entre un tercio y la mitad de una Raspberry. Estos dispositivos son compatibles y ofrecen características de hardware similares con algunas limitaciones que van desde una menor cantidad de memoria RAM, menos puertos USB hasta la ausencia de WiFi y bluetooth según el modelo.

Otro de los aspectos a mejorar en este dispositivo son los ambientes con muy baja intensidad de luz o nula, en este caso es posible evaluar la posibilidad de incorporar una cámara de luz infrarroja y adaptar este dispositivo para que pueda reconocer los gestos también en estos casos.



ANEXO A - PROTOCOLO DE CONTROL REMOTO INFRARROJO

Un tema central de esta tesina está compuesta por los componentes de hardware del dispositivo construido. Teniendo en cuenta esto vale la pena mencionar como funcionan sus partes. En este anexo se describirán como funciona el protocolo de control de señales infrarrojas para un control remoto de TV.

Las señales infrarrojas o IR, es radiación electromagnética (EMR) con longitudes de onda más largas que las de la luz visible, y por lo tanto es invisible. Se extiende desde el borde rojo nominal del espectro visible a 700 nanómetros (frecuencia 430 THz), hasta 1 mm (300 GHz) [18] (aunque los láseres especialmente pulsados pueden permitir que los humanos detecten la radiación IR hasta 1050 nm.).

Infrarrojo fue descubierto en 1800 por el astrónomo Sir William Herschel, quien descubrió un tipo de radiación invisible en el espectro con menor energía que la luz roja, por medio de su efecto sobre un termómetro. Se descubrió que algo más de la mitad de la energía total del Sol llegaba a la Tierra en forma de infrarrojos. El equilibrio entre la radiación infrarroja absorbida y emitida tiene un efecto crítico en el clima de la Tierra.

En la mayoría de los sistemas de transmisión de control remoto, solo se requieren pequeñas velocidades de datos para transmitir las funciones de control del equipo de entretenimiento doméstico. La fiabilidad de la transmisión es esencial ya que no se permite una interpretación incorrecta de un código transmitido. Las señales corruptas deben ser ignoradas. En la mayoría de los esquemas de codificación, los comandos se repiten hasta que el dispositivo controlado a distancia reacciona como se deseé. El

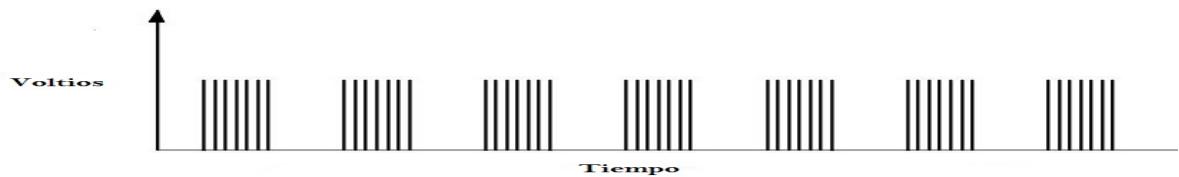


Figura A.1: Cada pulso se enciende y apaga a una frecuencia de 38 kHz

operador puede observar directamente el resultado de presionar una tecla mediante retroalimentación visual. Debido a que las señales IR están confinadas dentro de una habitación y debido a que solo hay un corto período de transmisión de datos con cada pulsación de tecla, no hay restricciones legales para la transmisión IR en la banda de frecuencia entre 30 kHz y 56 kHz.

Conceptos básicos de comunicación IR

La radiación IR es simplemente luz que no podemos ver, lo que la hace ideal para la comunicación. Las fuentes IR están a nuestro alrededor. El sol, las bombillas o cualquier otra cosa con calor es muy brillante en el espectro IR. Al utilizar un control remoto de un televisor, se usa un LED IR para transmitir información. Entonces, ¿cómo el receptor IR en su televisor selecciona las señales de su control remoto entre todos los infrarrojos? La respuesta es que la señal IR está modulada. La modulación de una señal es como asignar un patrón a sus datos, de modo que el receptor sepa escuchar.

Un esquema de modulación común para la comunicación IR es algo llamado modulación de 38 kHz. Hay muy pocas fuentes naturales que tengan la regularidad de una señal de 38 kHz, por lo que un transmisor IR que envíe datos a esa frecuencia se destacaría entre la IR ambiental. Los datos IR modulados a 38kHz son los más comunes, pero se pueden usar otras frecuencias (Figura A.1).

Cuando presiona una tecla en su control remoto, el LED IR transmisor parpadeará rápidamente por una fracción de segundo, transmitiendo datos codificados a su dispositivo.

Si se conectara un osciloscopio al LED IR de un control remoto de TV, se vería una señal similar a la figura A.1. Esta señal modulada es exactamente lo que ve el sistema receptor. Sin embargo, el objetivo del dispositivo receptor es demodular la señal y generar una forma de onda binaria que pueda ser leída por un microcontrolador. Al leer el pin OUT del TSOP382 con la onda desde arriba, se verá algo muestra la figura A.2.

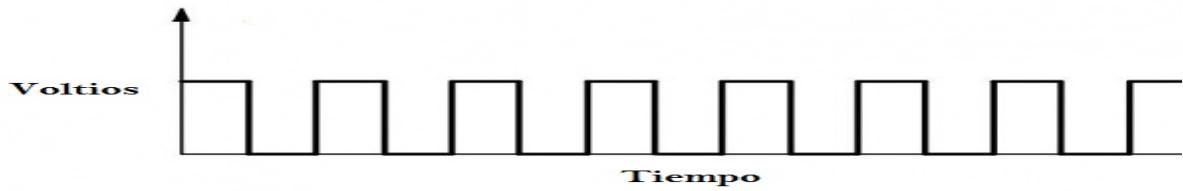


Figura A.2:

Controlando el espaciado entre las señales moduladas transmitidas, la forma de onda puede leerse mediante un pin de entrada en un microcontrolador y decodificarse como un flujo de bits en serie.

Protocolo Infrarrojo

El protocolo de transmisión IR utiliza codificación de distancia de pulso de los bits de los mensajes. Cada ráfaga de pulso (marca - RC transmisor ENCENDIDO) tiene una longitud de 562.5, a una frecuencia de transferencia de 38kHz. Los bits lógicos se transmiten como unos y ceros seguidos de espacios de tiempo.

Al transmitir o recibir códigos de control remoto utilizando el protocolo de transmisión IR, funciona de manera óptima cuando la frecuencia de transmisión (utilizada para la modulación / demodulación) se establece en 38.222 kHz.

Cuando se presiona una tecla en el control remoto, el mensaje transmitido consta de lo siguiente, en orden:

- Una ráfaga de pulso inicial de 9 ms (16 veces la longitud de la ráfaga de pulso utilizada para un bit de datos lógicos).
- Un espacio de 4.5ms.
- La dirección de 8 bits para el dispositivo receptor.
- La inversa lógica de 8 bits de la dirección.
- El comando de 8 bits.
- El inverso lógico de 8 bits del comando.
- Una ráfaga de pulso final de 562.5 para indicar el final de la transmisión del mensaje.

ANEXO A. ANEXO A - PROTOCOLO DE CONTROL REMOTO INFRARROJO

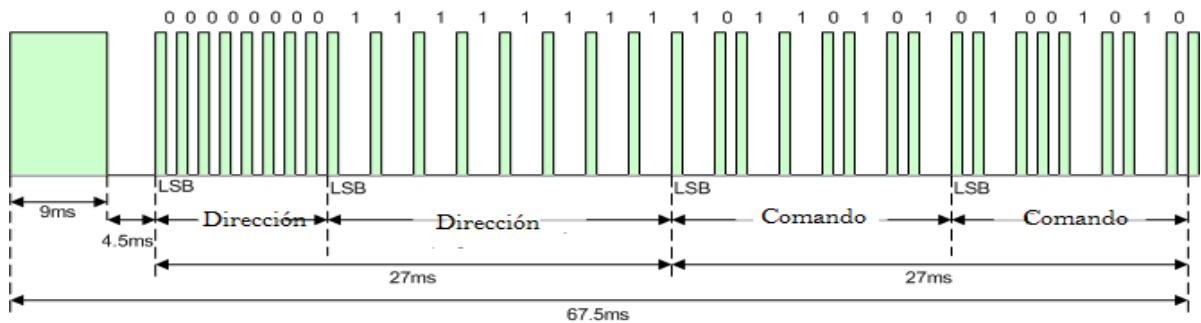


Figura A.3: Protocolo IR

A los cuatro bytes de bits de datos se les envía cada bit menos significativo primero. La Figura A.3 ilustra el formato de un cuadro de transmisión IR.

Paquetes para el desarrollo de software

Para desarrollar software que se pueda comunicar con los dispositivos que reciben o envían señales infrarrojas, existen librerías que permiten codificar y decodificar esas señales y utilizarlas con lenguajes de alto nivel. En el caso de esta tesina se utilizó la librería LIRC (Linux Infrared Remote Control), que está desarrollada para Linux y nos permite enviar y recibir señales infrarrojas, y además soporta una amplia variedad de hardware.

Lirc: es un paquete que le permite decodificar y enviar señales infrarrojas de muchos (pero no todos) controles remotos comunes. Los últimos kernels de linux hacen posible usar algunos controles remotos IR como dispositivos de entrada regulares. A veces esto hace que LIRC sea redundante. Sin embargo, LIRC ofrece más flexibilidad y funcionalidad y sigue siendo la herramienta adecuada en muchos escenarios.

La parte más importante de LIRC es el daemon lircd que decodifica las señales IR recibidas por los controladores de dispositivo y proporciona la información en un socket. También acepta comandos para que se envíen señales IR si el hardware lo admite.

BIBLIOGRAFÍA

- [1] *Gestigon gesture tracking - techcrunch disrupt*, October 2016.
- [2] T. A. S. ADAM KENDON AND J. UMIKER-SEBEOK, *Nonverbal Communication, Interaction, and Gesture: Selections from SEMIOTICA.*, 1981.
- [3] Z. H. AL-TAIRI, R. W. RAHMAT, M. I. SARIPAN, AND P. S. SULAIMAN, *Comparative study of skin color detection and segmentation in hsv and ycbcr color space*, Journal of Information Processing Systems, 10 (2014), pp. 283 – 299.
- [4] BMW, *The future of luxury is here @ONLINE*, 2017.
- [5] J. M. CHAVES-GONZÁLEZ, M. A. VEGA-RODRÍGUEZ, J. A. GÓMEZ-PULIDO, AND J. M. SÁNCHEZ-PÉREZ, *Detecting skin in face recognition systems: A colour spaces study*, Digit. Signal Process., 20 (2010), pp. 806–823.
- [6] M.-Y. CHEN, L. MUMMERT, P. PILLAI, A. HAUPTMANN, AND R. SUKTHANKAR, *Controlling your tv with gestures*, in Proceedings of the International Conference on Multimedia Information Retrieval, MIR '10, New York, NY, USA, 2010, ACM, pp. 405–408.
- [7] D. EFRON, *Gesture and environment*, 1941.
- [8] M. ENCYCLOPEDIA, *touchless user interface definition*, July 2017.
- [9] R. B. S. D. X.-F. M. C. K. K. E. M. F. QUEK, D. MCNEILL AND R. ANSARI, *Multimodal human discourse: gesture and speech*, 2002.
- [10] R. C. GONZALEZ, *Digital Image Processing, Third Edition*, 2008.
- [11] A. HASSANAT, M. AL-AWADI, E. BTOUSH, A. AL-BTOUSH, E. ALHASANAT, AND G. ALTARAWNEH, *New mobile phone and webcam hand images databases for personal authentication and identification*, Procedia Manufacturing, 3 (2015), pp. 4060 – 4067.

BIBLIOGRAFÍA

- 6th International Conference on Applied Human Factors and Ergonomics and the Affiliated Conferences, AHFE 2015.
- [12] R. J. HOWLETT, D.-L. DINH, J. T. KIM, AND T.-S. KIM, *Hand gesture recognition and interface via a depth imaging sensor for smart home appliances*, Energy Procedia, 62 (2014), pp. 576 – 582.
 - [13] P. KAKUMANU, S. MAKROGIANNIS, AND N. BOURBAKIS, *A survey of skin-color modeling and detection methods*, Pattern Recogn., 40 (2007), pp. 1106–1122.
 - [14] ——, *A survey of skin-color modeling and detection methods*, Pattern Recogn., 40 (2007), pp. 1106–1122.
 - [15] B. KANG, K. TAN, H. TAI, D. TRETTER, AND T. Q. NGUYEN, *Hand segmentation for hand-object interaction from depth map*, CoRR, abs/1603.02345 (2016).
 - [16] R. KHAN, A. HANBURY, J. STÖTTINGER, AND A. BAIS, *Color based skin classification*, Pattern Recogn. Lett., 33 (2012), pp. 157–163.
 - [17] K.-S. H. KUE-BUM LEE, JUNG-HYUN KIM, *An implementation of multi-modal game interface based on pdas*, 2007.
 - [18] S. C. LIEW, *Electromagnetic waves*, Centre for Remote Imaging, Sensing and Processing. Retrieved, (2006).
 - [19] R. LOCKTON, *Hand gesture recognition using computer vision*, (2002).
 - [20] M. LUCAS, *usens shows off new tracking sensors that aim to deliver richer experiences for mobile vr*, August 2016.
 - [21] A. MALIMA, E. OZGUR, AND M. CETIN, *A fast algorithm for vision-based hand gesture recognition for robot control*, in 2006 IEEE 14th Signal Processing and Communications Applications, April 2006, pp. 1–4.
 - [22] W. W. P. McCULLOCH, *A Logical Calculus of Ideas Immanent in Nervous Activity*, 1943.
 - [23] D. MCNEILL, *Hand and mind, what gestures reveal about thought*, 1992.
 - [24] T. PIUMSOMBOON, A. CLARK, M. BILLINGHURST, AND A. COCKBURN, *User-Defined Gestures for Augmented Reality*, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 282–299.

- [25] F. K. H. QUEK, *Toward a vision-based hand gesture interface*, August 1994.
- [26] F. QUIROGA, *Reconocimiento de gestos dinámicos*, 2014.
- [27] S. REIFINGER, F. WALLHOFF, M. ABLASSMEIER, T. POITSCHKE, AND G. RIGOLL, *Static and dynamic hand-gesture recognition for augmented reality applications*, in Proceedings of the 12th International Conference on Human-computer Interaction: Intelligent Multimodal Interaction Environments, HCI'07, Berlin, Heidelberg, 2007, Springer-Verlag, pp. 728–737.
- [28] D. L. REILLY, L. N. COOPER, AND C. ELBAUM, *A neural model for category learning*, Biological Cybernetics, 45 (1982), pp. 35–41.
- [29] Y. C. S. RICCI, P. E., *Comportamiento no verbal y comunicación*, 1980.
- [30] F. RONCHETTI, F. QUIROGA, C. ESTREBOU, L. LANZARINI, AND A. ROSETE, *Sign language recognition without frame-sequencing constraints: A proof of concept on the argentinian sign language*, in Advances in Artificial Intelligence - Lecture Notes in Computer Science, Springer International Publishing, 2016.
En prensa.
- [31] G. E. W.-R. J. RUMELHART, DAVID E.; HINTON, *Learning representations by back-propagating errors*, (1986).
- [32] R. V. SHABNAM KHAN, *Gesture recognition technology*, 2017.
- [33] A. SHIMADA, T. YAMASHITA, AND R. I. TANIGUCHI, *Hand gesture based tv control system - towards both user and machine-friendly gesture applications*, in Frontiers of Computer Vision, (FCV), 2013 19th Korea-Japan Joint Workshop on, Jan 2013, pp. 121–126.
- [34] D. C. SON LAM PHUNG, ABDESELAM BOUZERDOUM, *Skin segmentation using color pixel classification: Analysis and comparison*, Jun 2005.
- [35] A. SONI, D. K. LOBIYAL, K. B. SHAIK, P. GANESAN, V. KALIST, B. SATHISH, AND J. M. M. JENITHA, *3rd international conference on recent trends in computing 2015 (icrtc-2015) comparative study of skin color detection and segmentation in hsv and ycbcr color space*, Procedia Computer Science, 57 (2015), pp. 41 – 48.

BIBLIOGRAFÍA

- [36] C. SUI, N. M. KWOK, AND T. REN, *A restricted coulomb energy (rce) neural network system for hand image segmentation*, in Computer and Robot Vision (CRV), 2011 Canadian Conference on, May 2011, pp. 270–277.
- [37] D. TERDIMAN, *Leap motion: 3d hands-free motion control, unbound*, Pattern Recogn., (2012).
- [38] C. B.-S. B. THOMAS G. ZIMMERMAN, JARON LANIER AND Y. HARVILL, *A hand gesture interface device*, 1987.
- [39] F. VAFAEI, *Taxonomy of gestures in human computer interaction*, Aug 2013.
- [40] V. VEZHNEVETS, V. SAZONOV, AND A. ANDREEVA, *A survey on pixel-based skin color detection techniques*, in IN PROC. GRAPHICON-2003, 2003, pp. 85–92.
- [41] V. VEZHNEVETS, V. SAZONOV, AND A. ANDREEVA, *A survey on pixel-based skin color detection techniques*, in GraphiCon, 2003, pp. 85–92.
- [42] T. S. H. VLADIMIR I. PAVLOVIC, RAJEEV SHARMA, *Visual interpretation of hand gestures for human-computer interaction*, 1997.
- [43] A. WEXELBLAT, *An approach to natural gesture in virtual environments*, 1995.
- [44] ——, *Research challenges in gesture: Open issues and unsolved problems*, 1998.
- [45] Z. XU AND M. ZHU, *Color-based skin detection: survey and evaluation*, in 2006 12th International Multi-Media Modelling Conference, 2006, pp. 10 pp.–.
- [46] E. P. Y FRIESEN W. V., *The repertorie of non verbal behavior*, 1969.
- [47] E. P. Y FRIESEN W. V., *Hand movement. Journal of Communication*, 1972.
- [48] Y. J. YANG LIU, *A robust hand tracking and gesture recognition method for wearable visual interfaces and its applications*, 2004.
- [49] B. YEGNANARAYANA, *Artificial Neural Network*, 2006.
- [50] X. YIN, D. GUO, AND M. XIE, *Hand image segmentation using color and {RCE} neural network*, Robotics and Autonomous Systems, 34 (2001), pp. 235 – 250.
- [51] C. ZHI-HUA, K. JUNG-TAE, L. JIANNING, Z. JING, AND Y. YU-BO, *Real-time hand gesture recognition using finger segmentation*, The Scientific World Journal, (2014).