Dinámica basada en esqueleto y profundidad CNN + RNN Reconocimiento de gestos con las manos

Kenneth Lai y Svetlana N. Yanushkevich

Laboratorio de Tecnologías Biométricas, Departamento de Ingeniería Eléctrica e Informática Universidad de Calgary, Alberta, Canadá Correo electrónico:kelai@ucalgary.ca, syanshk@ucalgary.ca

Resumen—La actividad humana y el reconocimiento de gestos son un componente importante del dominio de la inteligencia ambiental en rápido crecimiento, en particular para ayudar a la vida y los hogares inteligentes. En este artículo, proponemos combinar el poder de dos técnicas de aprendizaje profundo, las redes neuronales convolucionales (CNN) y las redes neuronales recurrentes (RNN), para el reconocimiento automatizado de gestos manuales utilizando datos tanto de profundidad como de esqueleto. Cada uno de estos tipos de datos se puede utilizar por separado para entrenar redes neuronales para que reconozcan los gestos de las manos. Si bien se informó anteriormente que RNN funciona bien en el reconocimiento de secuencias de movimiento para cada articulación del esqueleto dada solo la información del esqueleto, este estudio tiene como objetivo utilizar datos de profundidad y aplicar CNN para extraer información espacial importante de las imágenes de profundidad. Juntos, el tándem CNN+RNN es capaz de reconocer una secuencia de gestos con mayor precisión. Además, se estudian varios tipos de fusión para combinar tanto la información del esqueleto como la de profundidad para extraer información temporal-espacial. Se logra una precisión general del 85,46 % en el conjunto de datos de gestos manuales dinámicos: 14/28.

Palabras clave: biometría, reconocimiento de gestos, redes neuronales convolucionales, redes neuronales recurrentes.

I. INTRODUCCIÓN

Este artículo se ocupa de las técnicas de inteligencia computacional ambiental basadas en biometría. Específicamente, nos centramos en la actividad humana y el reconocimiento de gestos en el contexto de la monitorización ambiental en hogares inteligentes, residencias asistidas o instalaciones sanitarias. En una casa inteligente, los sensores se pueden programar para conocer las rutinas diarias normales de un residente, que luego se pueden usar para realizar monitoreo y evaluación automatizados de la salud ambiental [1]. Por ejemplo, Pavel et al. [2] sugirió que existe una relación entre los patrones de movilidad y la capacidad cognitiva. La teoría se examinó observando los cambios en la movilidad y se encontró evidencia que respalda la relación entre la movilidad y la capacidad cognitiva. En la actual población que envejece, "los desafíos de mantener la movilidad y la función cognitiva hacen que sea cada vez más difícil seguir viviendo solo, lo que obliga a muchas personas a buscar residencia en instituciones clínicas" [3].

Lee y Dey [4] diseñaron un sistema de detección integrado para determinar si el residente adquiere más conciencia sobre sus capacidades funcionales cuando se le proporciona información sobre sus movimientos. La capacidad de realizar una evaluación automatizada de la calidad de las tareas y la salud cognitiva ha mejorado enormemente la precisión [5], [6]. Estas técnicas indican que se puede extraer información específica de un sensor y utilizarla para etiquetar la actividad realizada. Por ejemplo, algunas actividades como lavar los platos, tomar medicamentos y usar el teléfono se caracterizan por la interacción con objetos únicos.

El principal objetivo de este artículo es implementar un marco para el reconocimiento de actividades, incluido el reconocimiento de gestos. El reconocimiento de actividad tradicional utiliza principalmente imágenes RGB para el análisis. Los métodos actuales incorporan diferentes tipos de información, incluida la profundidad, el infrarrojo y el tiempo [7]. El método propuesto se centra en la creación de un marco que utilice información tanto profunda como esquelética en la tarea de reconocimiento de actividad/gesto manual. Para probar tal punto, hemos seleccionado la tarea específica de reconocer gestos dinámicos con las manos utilizando información de profundidad y esqueleto. Aplicamos las técnicas de aprendizaje profundo de última generación, como la red neuronal convolucional (CNN) y las redes neuronales recurrentes (RNN).

El artículo está estructurado de la siguiente manera: los trabajos relacionados se presentan en la Sección II, el marco del método propuesto en la Sección III, el diseño de experimentos y resultados experimentales en la Sección IV, y las conclusiones en la Sección V.

II. OBRAS RELACIONADAS

El reconocimiento de actividades y gestos es un ámbito que se investiga activamente, especialmente a la luz del desarrollo reciente de tipos nuevos y avanzados de sensores que recopilan datos múltiples y más precisos. En la tarea de reconocer gestos y actividades se han examinado diferentes espectros de datos, incluidos color, profundidad, GPS, aceleración, infrarrojos, etc.

Antes de la popularización de los sensores de profundidad económicos, los datos disponibles para el reconocimiento de gestos eran principalmente colores (intensidad de píxeles) o RGB. Uno de los métodos más comunes era usar marcadores de colores que indicaran diferentes regiones de la mano. Iwai et al. utilizó un guante de color en combinación con un árbol de decisión para realizar el reconocimiento de gestos [8]. Otro enfoque basado en el color de Bretzner et al. utilizó características de color de múltiples escalas [9].

Con el desarrollo de Microsoft Kinect, LeapMotion e Intel RealSense, la información 3D y de profundidad ahora se adquiere más fácilmente para su análisis. En [10] se propone un enfoque basado en características de profundidad extraídas que permiten crear una silueta de profundidad. En [11] se sugiere otro método que utiliza imágenes de profundidad para entrenar un bosque aleatorio de múltiples capas.

Motivados por la relación entre los gestos de las manos y el reconocimiento del lenguaje de señas, [12] propusieron un método para reconocer las formas de las manos utilizando bosques aleatorios aplicados a imágenes tanto en profundidad como en color. Otro enfoque utiliza profundidad y

Identificador de objetos digitales 10.1109/ICPR.2018.8545718
978-1-5386-3788-3/18/$31.00 c IEEE 2018

imágenes en color y varios tipos de descriptores espaciotemporales, combinados con diferentes opciones de núcleo para que la máquina de vectores de soporte encuentre la combinación óptima de reconocimiento [13].

En los últimos años, las técnicas de aprendizaje profundo han revolucionado el reconocimiento de patrones en general. Una CNN 3D [14] combina el aumento de datos espaciotemporales para realizar el reconocimiento de gestos en imágenes de profundidad y color. Ha logrado una tasa de clasificación del 77,5% en el conjunto de datos del desafío VIVA. Un método propuesto por Nagi et al. [15], sugieren el uso de "redes neuronales grandes y profundas de última generación que combinan convolución y agrupación máxima" para la extracción y clasificación de características. En [16], RNN se utiliza para modelar la información temporal en una secuencia de movimiento de la articulación del esqueleto para realizar el reconocimiento de gestos. De manera similar, [17] combina CNN con memoria a corto plazo (LSTM) para reconocer gestos dinámicos con las manos utilizando solo información basada en esqueletos.

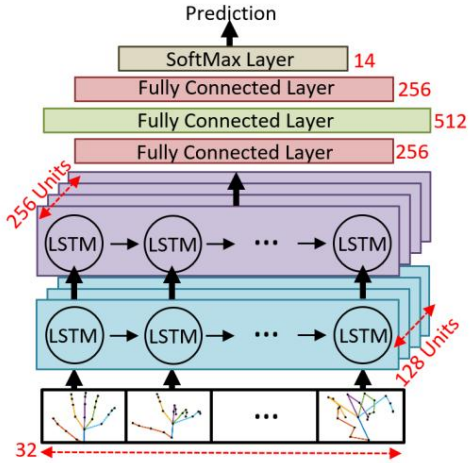


Fig. 1. La estructura de la red LSTM basada en esqueleto.

III. ESTRUCTURA

Esta sección describe el enfoque propuesto de utilizar puntos de profundidad y de esqueleto para reconocer un gesto con la mano.
El sistema consta de dos componentes principales: un CNN+RNN basado en profundidad (Fig. 3) y un RNN basado en esqueleto (Fig. 1).

El primer componente del sistema, CNN, está diseñado para extraer características de la imagen de profundidad de un sujeto que realiza un gesto con la mano. Sin embargo, dado que el procesamiento del reconocimiento dinámico de gestos depende de una secuencia ordenada de imágenes, se propone un RNN para complementar la CNN para extraer patrones temporales. La arquitectura general de la CNN y la red de memoria a corto plazo (LSTM) se muestra en la Fig. 3.
La estructura consta de tres componentes: la extracción de características basada en profundidad a través de CNN, el procesamiento de series de tiempo a través de RNN y la clasificación mediante un perceptrón multicapa (MLP). El primer componente, CNN, incluye seis capas convolucionales de 3x3 con una capa de agrupación máxima de 2x2 entre cada dos capas convolucionales. El segundo componente, RNN, incluye dos capas LSTM, cada una de las cuales consta de 256 unidades LSTM. El componente final, MLP, contiene tres capas completamente conectadas (FC) que constan de 256, 256 y 14 unidades, respectivamente.

El segundo componente del sistema incluye RNN, que extrae patrones temporales de los movimientos de puntos del esqueleto dentro de una secuencia. La estructura RNN se muestra en la Fig. 1 y es similar a la descrita en el primer componente, excepto por una capa FC adicional y varias unidades LSTM dentro de cada capa.

En general, el marco propuesto consta de dos componentes de red que utilizan información de esqueleto y profundidad para el reconocimiento de gestos de forma independiente. Dado que cada red es capaz de predecir un gesto seleccionado basándose en el tipo de información seleccionado, se espera que una fusión de ambas redes produzca un mayor rendimiento que sea capaz de seleccionar las mejores características tanto del esqueleto como del espacio espacial basado en profundidad. información y los patrones temporales inherentes entre una colección de fotogramas.

Hay diferentes formas de realizar la fusión como se ilustra en la Fig. 2. Las tres técnicas principales consideradas en este estudio incluyen la fusión a nivel de característica, la fusión a nivel de puntuación y la fusión a nivel de decisión.

La fusión a nivel de entidad se puede realizar en cualquier capa antes del MLP (que consta de las capas totalmente conectada, soft-max y de clasificación). En general, las capas de convolución en CNN y las capas LSTM en RNN son la parte de la red diseñada para extraer características de los datos de entrada. Se puede realizar una fusión a nivel de característica mediante una fusión de los resultados después de que los datos de entrada hayan pasado a través de una serie de capas de convolución/LSTM. Después de la fusión, se puede conectar un clasificador como MLP o una máquina de vectores de soporte para crear una red de fusión a nivel de características.

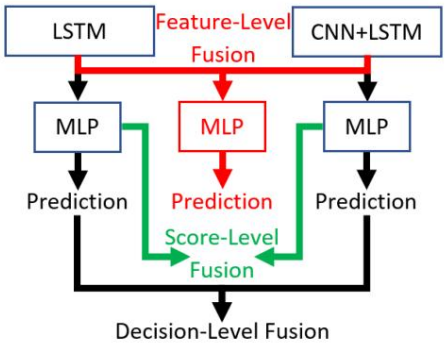


Fig. 2. Los tres niveles de fusión para las redes LSTM y CNN+LSTM.

De manera similar a la fusión a nivel de característica, la fusión a nivel de puntuación se puede realizar después o entre las capas completamente conectadas y soft-max. En este proceso, asumimos que las características se extrajeron con éxito de los datos de entrada y se pasaron a través de un clasificador seleccionado, lo que resultó en alguna puntuación o probabilidad. Dado que la puntuación está fuertemente correlacionada con la predicción de la red, la combinación de múltiples puntuaciones de diferentes redes proporcionará una predicción más confiable.

La fusión a nivel de decisión es comparable a una fusión a nivel de puntuación excepto que la fusión se realiza después de que la red

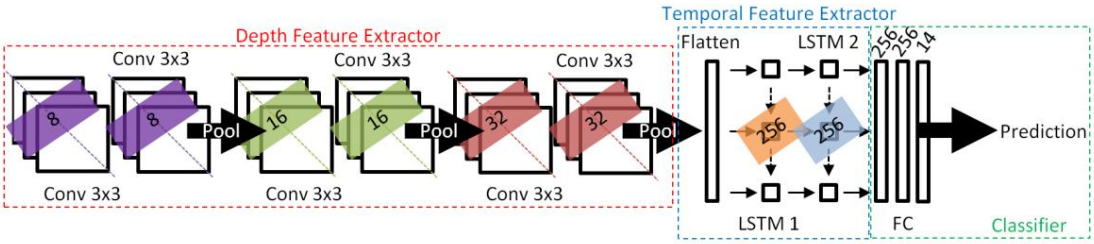


Fig. 3. La estructura de la red CNN+LSTM basada en profundidad.

predicción. Este tipo de fusión se basa completamente en el resultado previsto de la red y no está asociado con la puntuación/probabilidad utilizada para la decisión. La salida prevista de una red se define en función de la salida de probabilidad/ puntuación de la capa soft-max. Los dos métodos más generales para que una red genere una predicción se basan en un umbral de decisión o en el orden de clasificación. El método del umbral de decisión se basa en seleccionar un valor arbitrario para cada red. Cualquier puntuación superior al valor seleccionado se considera aceptada, de lo contrario se rechaza. En cuanto al método de orden de clasificación, todas las puntuaciones se agrupan y ordenan de modo que cuanto más alta sea la lista, mayor será la probabilidad de aceptación. Un uso general es el reconocimiento de rango 1, que solo considera la predicción superior como la salida prevista de la red.

En este artículo, hemos explorado la fusión tanto a nivel de característica como a nivel de puntuación, pero hemos excluido la fusión a nivel de decisión. La fusión a nivel de decisión no se examina en este artículo porque está correlacionada con la fusión a nivel de puntuación y solo hay dos decisiones de red para combinar.

Además de los dos niveles de fusión, existen varios tipos de fusión de los cuales consideramos concatenación, promediación y máxima. En este estudio se realizó fusión mediante concatenación a nivel de características porque genera un nuevo conjunto de características que considera tanto la profundidad extraída como la información del esqueleto. La fusión que utiliza promedio y máximo solo se aplica a la fusión a nivel de puntuación porque cada puntuación representa la fortaleza de predicción de una red. Se espera que el método de promedio genere una puntuación más confiable, ya que se basa en dos tipos de información y redes. Finalmente, la técnica máxima pondrá más énfasis en la capacidad de cada red para predecir gestos específicos.

IV. EXPERIMENTOS

El experimento se lleva a cabo en cada componente del marco propuesto que se muestra en las Fig. 3 y 1 de forma independiente.
Seguimos la misma configuración experimental indicada en [18] que utilizó una estrategia de validación cruzada de dejar un sujeto fuera.
Con base en esta estrategia, cada red propuesta se entrena en 19 sujetos y se prueba en el sujeto restante, lo que da como resultado una validación cruzada de 20 iteraciones.

La red CNN basada en profundidad se entrena durante 20 épocas utilizando un tamaño de lote mínimo de 32 con el optimizador Adadelta [19] con los parámetros predeterminados de lr = 1,0, ρ = 0,95 y = 1e −08. La entrada de la red se basa en el recorte. Imágenes de mano del conjunto de datos DHG-14/28 redimensionadas a 227x227. Se eligió una cantidad baja de épocas porque los pesos están diseñados para inicializar los pesos en la red CNN+LSTM, por lo que no es necesario encontrar la mejor solución óptima.

De manera similar, CNN-LSTM basado en profundidad utiliza el optimizador Adadelta con parámetros predeterminados con un tamaño de imagen de entrada de 227x227. La red se entrena con un tamaño de minilote de 16, un paso de tiempo de 32 y en 100 épocas. Se seleccionó un paso de tiempo de 32 porque el número de imágenes clave para un gesto varía entre 7 y 149 con un promedio de 34,59 fotogramas clave.

Para la red LSTM basada en esqueleto, se utiliza el optimizador Adam [20] con parámetros predeterminados de lr = 0,001, β1 = 0,9, β2 = 0,999 y = 1e −08 para entrenar la red.
Se seleccionó un paso de tiempo de 32 y un tamaño de mini lote de 32 porque corresponde al número de fotogramas clave promedio para cada gesto. Tenga en cuenta que los datos de entrada están representados por los puntos de coordenadas 2D que indican las articulaciones del esqueleto de una mano disponibles en el conjunto de datos DHG-14/28.

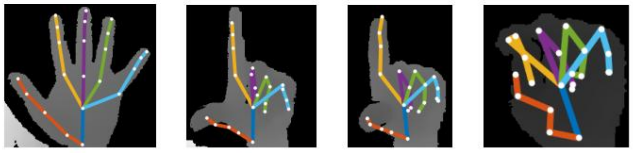
Para las redes de fusión a nivel de puntuación y a nivel de características, se utiliza el optimizador Adadelta con parámetros predeterminados. Los datos de entrada constan de secuencias de imágenes de profundidad (227x227) y secuencias de ubicaciones de articulaciones esqueléticas 2D (44x1). El entrenamiento se ejecuta durante 100 épocas con un paso de tiempo de 32 y un tamaño de minilote de 16.

A. Conjuntos de datos

Se eligió como base de datos el gesto dinámico de la mano 14/28 (DHG-14/28) [18], y esta es una de las pocas bases de datos que contiene datos recopilados utilizando un sensor de cámara de profundidad (Intel RealSense F200). Se encuentra disponible información de profundidad y esqueleto para varios gestos con las manos. En el conjunto de datos DHG-14/28 [18], hay 20 individuos únicos que realizan 5 iteraciones de 14 gestos utilizando dos tipos de configuraciones de dedos, formando así 28 conjuntos de gestos, para un total de 2800 secuencias.

La información de profundidad se guarda en forma de imágenes con una resolución de 480x640 en 16 bits. La información del esqueleto contiene 22 ubicaciones de articulaciones de una mano descritas en coordenadas 2D y 3D guardadas en formato vectorial 44x1 y 66x1, respectivamente.

Para el conjunto de datos DHG-14/28, cada gesto se clasifica individualmente en dos categorías principales: gestos detallados y gruesos. La Tabla I proporciona una lista de todos los gestos y las categorías de grano correspondientes.



(a) (b) (C) (d)

Fig. 4. Ilustraciones del gesto de agarre del DHG-14/28 [18]. (a) Marco 1, (b) Marco 40, (c) Marco 50, (d) Marco 60.

TABLA I

LISTA DE GESTOS EN EL CONJUNTO DE DATOS DHG-14/28

Gesto	Grano	Nombre de etiqueta
Agarrar	Bien G	
Grifo	T gruesa	
Expandir	Bien	mi
Pellizco	Bien	PAG
Rotación en el sentido de las agujas del reloj	Bien	R-CW
Rotación en sentido antihorario	Bien	R-CCW
Desliza a la derecha	SR grueso	
Deslizar a la izquierda	SL grueso	
Muévase hacia arriba	SU grueso	
Deslizar hacia abajo	SD gruesa	
Desliza X	SX grueso	
Deslizar V	SV grueso	
Desliza +	Grueso S-+	
Agitar	Sh grueso	

B. Preprocesamiento

Las entradas de datos para la red CNN+LSTM son las imágenes de profundidad. Las imágenes de profundidad contienen originalmente información de 16 bits por píxel y se normalizaron para que el valor de píxel oscile entre 0 y 1. Además, las imágenes de manos se recortan de todo el fotograma según la región de interés proporcionada por el conjunto de datos. Por último, para cada secuencia, solo se utilizan para el reconocimiento los fotogramas entre el movimiento inicial y final del gesto.

En el caso de la red LSTM, solo se utilizan puntos de esqueleto 2D para el procesamiento. Para un gesto seleccionado, cada punto del esqueleto en una secuencia se normaliza restando cada punto por la ubicación de la palma del cuadro inicial. Además, para el reconocimiento sólo se utilizan las secuencias marcadas entre el inicio y el final del gesto.

C. Resultados experimentales

Para comparar el rendimiento del método propuesto, incluimos las tasas de reconocimiento de otros métodos que también utilizaron la misma base de datos para experimentos. La Tabla II ilustra 14 gestos únicos y las tasas de reconocimiento de métodos seleccionados examinados utilizando el conjunto de datos DHG-14/28. Como se menciona en [16], no es suficiente proporcionar únicamente las tasas de clasificación promedio. Por lo tanto, a modo de comparación, en la Tabla II se proporcionan las tasas de clasificación mejor, peor y promedio para cada categoría de grano de gesto. Además, los resultados de la Tabla II se examinan más a fondo en función del grano del gesto (fino, grueso y/o ambos tipos de grano) como se categorizó previamente en la Tabla I.

El estudio experimental realizado muestra que tanto el método propuesto de utilizar CNN+LSTM basado en profundidad como la red LSTM basada en esqueleto demuestran un rendimiento relativamente similar. Las filas 4 a 6 de la Tabla II representan el rendimiento de las redes de fusión propuestas. FL-fusion-Concat muestra las tasas de reconocimiento logradas por la fusión a nivel de características mediante la concatenación de las características extraídas de la red LSTM basada en esqueleto y la red CNN+LSTM basada en profundidad.
SL-fusion-Average informa el rendimiento obtenido por la fusión a nivel de puntuación promediando los resultados de las capas soft-max de cada uno de los esqueletos y las redes basadas en profundidad.
SL-fusion-Maximum representa la fusión a nivel de puntuación que predice el resultado basándose en encontrar las puntuaciones máximas entre el esqueleto y la red profunda.

De las tres redes de fusión, SL-fusion-Average tiene el mejor rendimiento, mientras que FL-fusion-Concat tiene el peor rendimiento. De la Tabla II, FL-fusion-Concat (fila 4) tiene un rendimiento peor que la red Skeleton LSTM predeterminada, lo que indica que el proceso de fusión de información de profundidad y esqueleto a nivel de característica degrada el rendimiento general. Cabe señalar que, aunque el rendimiento general se reduce, la tasa de reconocimiento del gesto detallado basado en la profundidad es un 3,60 % mayor que la del gesto detallado basado en el esqueleto. SL-fusion-Average (fila 5) y SL-fusion-Maximum (fila 6) muestran un rendimiento similar; SL-fusion-Average funciona un 0,1% mejor para ambos tipos de gestos granulados. El rendimiento de la fusión a nivel de puntuación muestra que este nivel de fusión da como resultado un mejor rendimiento en comparación con las redes de esqueleto y profundidad independientes.
Aunque la red esquelética proporciona un mayor rendimiento de forma independiente, la puntuación combinada entre ambas redes produce el mejor rendimiento porque la profundidad proporciona información que puede perderse en el proceso de extracción de las uniones esqueléticas.

Para examinar más a fondo el rendimiento del reconocimiento de gestos en términos de tasas de clasificación promedio, se creó una matriz de confusión (Fig. 6 y 5). Ilustra la precisión de la clasificación realizada por el método propuesto cuando se utiliza la fusión a nivel de puntuación de información de esqueleto y profundidad para 14 y 28 gestos, respectivamente. Para cada figura, el eje x representa la predicción de la red, mientras que el eje y representa el gesto real. Por ejemplo, el gesto R-CW (quinta fila) indica que la red es capaz de predecir correctamente el 79,5% de todos los gestos R-CW, mientras identifica erróneamente el otro 20,5% como otros gestos. La figura 6 muestra que el mayor fallo de la red de fusión se produce en la predicción del gesto de agarrar (primera fila), que se identifica erróneamente el 67,0 % del tiempo, del cual el 42,0 % se clasifica como gesto de pellizco. [16] y [18] han observado que este gesto de agarre específico es difícil de distinguir del movimiento de pellizco. También se puede ver una observación similar en la Fig. 5.

Según las tasas de reconocimiento de gestos de las figuras 6 y 5, la pérdida de precisión al eliminar la diferenciación de los dedos (sugerida en [18]) se calcula en 0,07286, que es siete veces mayor que el 0,0114 obtenido en [18]. En promedio, la fusión proporcionó una tasa de reconocimiento del 85,46% y 74,19% para 14 y 28 gestos, respectivamente. La transición entre 14 y 28 gestos supone una disminución del 11,27% del cual el 7,29% proviene de confusión intragestual. Esto indica que la red de fusión aún no está optimizada para los 28 gestos y, al utilizar los pesos de la red previamente entrenados en 14 gestos, el

TABLA II

TASAS DE RECONOCIMIENTO (%) DEL CONJUNTO DE DATOS DHG-14

Método	Bien			Grueso			Ambos		
	Lo mejor	de lo peor	Promedio ±	Mejor	Peor	Promedio ± Estándar	Lo mejor	de lo peor	Promedio ± Estándar
Profundidad CNN	52,89	24,83	estándar 37,05 ± 7,89	38,92	20,61	73,50 29,57±4,51	40,19	25,05	32,24 ± 4,64
Profundidad CNN + LSTM	90,00	52,00	± 10,93 92,22	58,89	69,90 ± 9,91	96,67 77,06 ± 8,73	85,00	58,57	75,79 ± 7,23
Esqueleto LSTM	82,00	54,00	76,67	72,90 ± 10,30	98,89	78,89 76,00 ± 89,00 ± 5,40	91,43	71,43	82,18 ± 5,32
FL-fusion-Concat	90.00	48.00	10,51	81,11	75,30 ± 10,89	98,89 78,89 86,83 ± 4,68	87,86	67,86	81,86 ± 5,38
SL-fusión-promedio	92,00	52,00	61,20 ± 12,37	97,78	74,67	78,50 ± 11,44 90,72 ± 4,64	95,00	90,94 ± 72,86	85,46 ± 5,16
SL-fusión-Máximo	94,00	54,00	96,67	64,44	76,90 ± 9,19	97,78 72,22 4,36	91,43	86,44± 7,94	93,57 71,43 85,36 ±5,06
Esqueleto [16]	86.00	42.00	78,00	N/ AN/A	73,60	N/A/A 66,90 81,94 ± 8,17	90,00	89,00 ± 67,86	77,43 ± 6,82
Función de movimiento [16]	84,00	46,00				94 .29	89,83	No aplicable 58,57	78,21 ± 7,49
Función esqueleto + movimiento [16]	90,00	56,00				88,33	No aplicable	85,94 67,86	84,68 ± 6,67
Esqueleto 3D CNN+LSTM [17]	N/AN/A							No	85,61
Basado en esqueletos, De Smedt [18]	N/AN/A							disponible	83.07
Basado en profundidad, De Smedt [18]	N/AN/A			N/AN/A			N / A	N / A	79,14

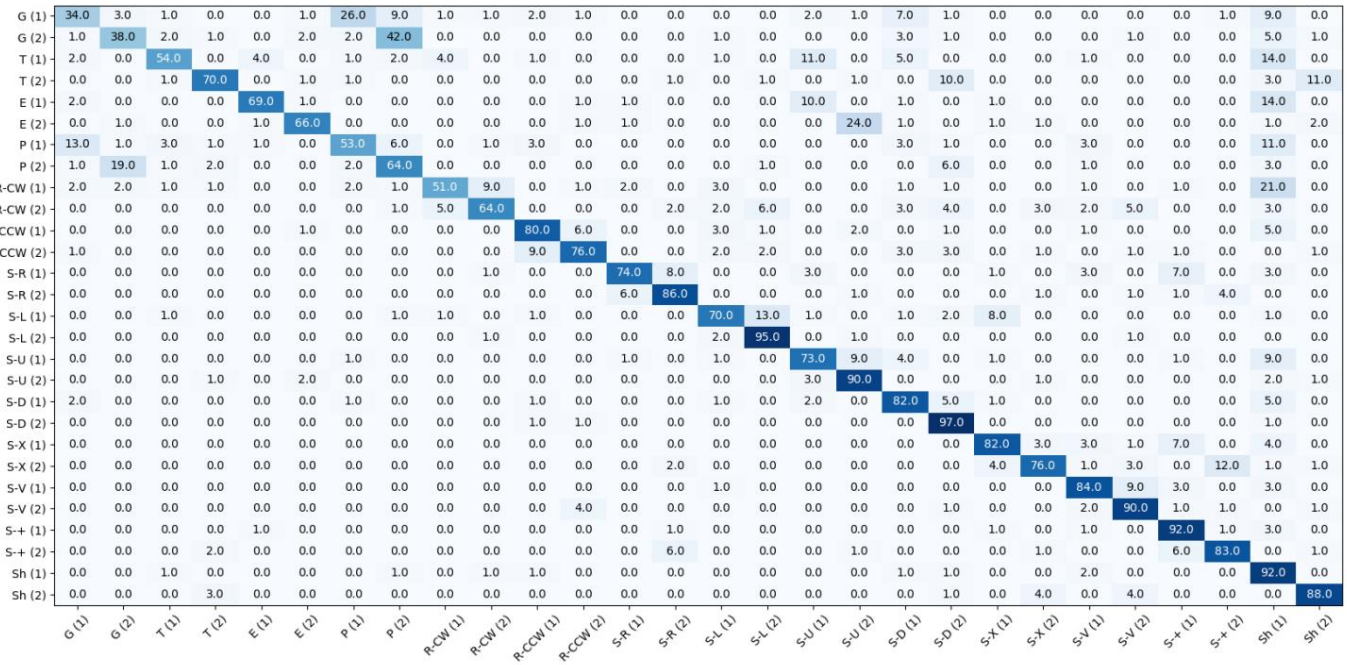


Fig. 5. La matriz de confusión que muestra la precisión del reconocimiento de gestos cuando se utiliza la fusión del nivel de puntuación (promedio) de la información del esqueleto y de la profundidad. en 28 gestos.

La red general produce un mayor error.

V. CONCLUSIONES

Este artículo contribuye al estudio del monitoreo ambiental basado en sensores de la actividad humana y el reconocimiento de gestos. Específicamente, se centra en la tarea de gesto dinámico con la mano. reconocimiento, utilizando datos tanto del esqueleto como de la profundidad adquiridos por sensores RGB de profundidad y potentes técnicas de aprendizaje profundo. Los resultados del estudio experimental realizado son resumido de la siguiente manera:

1) Uso de datos de profundidad, además del espectro visual, RGB datos, tiene potencial para mejorar enormemente el rendimiento de algoritmos de reconocimiento de gestos con las manos que se basan en patrones temporales. Mientras que el uso de sólo datos de profundidad con CNN conduce a una tasa de reconocimiento promedio del 32,24%. Sin embargo, utilizar CNN para la extracción de características y Al pasar esas características al RNN, se muestra que el rendimiento general aumenta hasta un 75,79%. Esto resulta indican que CNN por sí sola es insuficiente para extraer suficiente información de una imagen de profundidad singular para predecir correctamente el gesto deseado. Combinando la CNN con la RNN, la nueva red es capaz de reconocer patrones a partir de las características extraídas de un CNN a lo largo de múltiples fotogramas. Otro aspecto que influye en el rendimiento es el parámetro de paso de tiempo elegido. que dependiendo de los datos de entrada puede ser severamente truncado cuando se le dan secuencias largas o se rellena con información en blanco para secuencias cortas.

2) El rendimiento del uso de datos esqueléticos ha mostrado un tasa de reconocimiento del 82,18%, que es más alta en comparación con las redes basadas en profundidad. Sin embargo, al examinar la tasas de reconocimiento para cada gesto de grano, el basado en profundidad Este enfoque produce una tasa de reconocimiento más alta para las personas más finas.

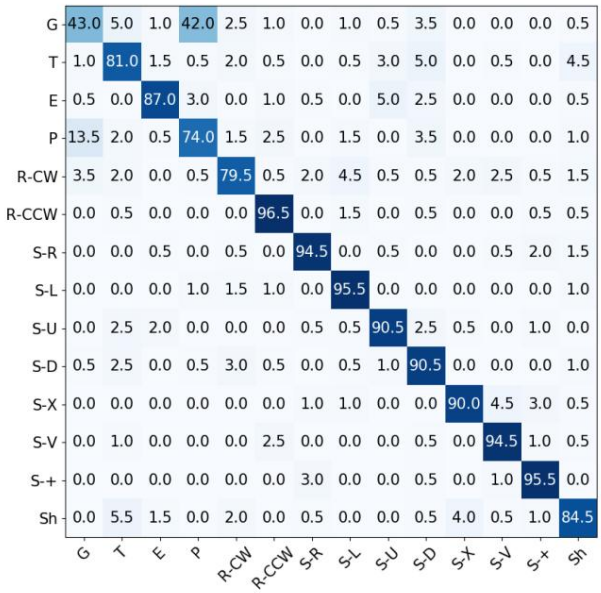


Fig. 6. La matriz de confusión que muestra la precisión del reconocimiento de gestos cuando se utiliza la fusión del nivel de puntuación (promedio) de la información esquelética y profunda en 14 gestos.

gesto granulado. Según esta observación, la fusión de redes basadas en profundidad y en esqueleto debería producir una tasa de reconocimiento general más alta. Los experimentos muestran que al utilizar una fusión de nivel de puntuación, se logra una tasa de reconocimiento del 85,46%, que es superior a las tasas del 82,18% y 76,50% para los enfoques de esqueleto independiente y de profundidad, respectivamente. Además de las tasas generales de reconocimiento más altas, las tasas para cada tipo de gesto de grano también aumentaron. Para los gestos detallados, mientras que la tasa para los enfoques basados en profundidad y esqueleto obtenidos por separado son del 73,50 % y 67,20 %, respectivamente, la tasa combinada es del 76,00 %. Para los gestos de grano grueso, los basados en profundidad (78,17 %) y los basados en esqueleto (89,00 %), cuando se combinan, muestran una tasa del 90,72 %.

3) La fusión aplicada permitió lograr muy buen rendimiento para el conjunto de datos que consta de 14 gestos. Cuando se aplica el mismo enfoque a la versión de 28 gestos, el rendimiento se degrada al 74,19%, lo que supone una disminución del ≈11%. La razón principal de esta disminución es que los pesos de las redes están optimizados para reconocer 14 gestos. Por lo tanto, se puede lograr un mejor rendimiento entrenando cada red de forma independiente para reconocer 28 gestos antes de la fusión.

Como trabajo futuro, anticipamos ampliar el marco propuesto hacia el reconocimiento de la actividad humana que se aplicaría a diversos escenarios de vida asistida, atención médica y interacción hombre-máquina.

EXPRESIONES DE GRATITUD

Este proyecto fue parcialmente apoyado por el Consejo de Investigación de Ingeniería y Ciencias Naturales de Canadá (NSERC) a través de Discovery. Beca “Interfaces Inteligentes Biométricas”; la Beca Reina Isabel II de la Provincia de Alberta y la Universidad de Calgary W21C (subvención de VPR “Innovaciones en atención médica domiciliaria para apoyar a una población que envejece”).

REFERENCIAS

[1] PN Dawadi, DJ Cook y M. Schmitter-Edgecombe, “Evaluación automatizada de la salud cognitiva mediante el monitoreo inteligente de tareas complejas en el hogar”, IEEE Trans. sobre sistemas, hombre y cibernética: sistemas, vol. 43, núm. 6, págs. 1302-1313, 2013.

[2] M. Pavel, A. Adami, M. Morris, J. Lundell, T. Hayes, H. Jimison y J. Kaye, “Evaluación de la movilidad mediante respuestas relacionadas con eventos”, en Transdisciplinary Conf. sobre diagnóstico distribuido y atención médica domiciliaria, 2006, págs. 71–74.

[3] A. Arcelus, MH Jones, R. Goubran y F. Knoefel, “Integración de tecnologías domésticas inteligentes en un sistema de seguimiento de la salud para personas mayores”, en Int. Conf. sobre talleres de aplicaciones y redes de información avanzadas, vol. 2, 2007, págs. 820–825.

[4] ML Lee y AK Dey, “Evaluación integrada del envejecimiento de adultos: una validación de concepto con las partes interesadas”, en Int. Conf. sobre tecnologías informáticas generalizadas para la atención sanitaria, 2010, págs. 1–8.

[5] DJ Cook, “Aprendizaje de modelos de actividad generalizada para espacios inteligentes”, IEEE Intelligent Systems, vol. 2010, núm. 99, pág. 1, 2010.

[6] E. Kim, S. Helal y D. Cook, “Reconocimiento de actividad humana y descubrimiento de patrones”, IEEE Pervasive Computing, vol. 9, núm. 1, págs. 48–53, 2010.

[7] HS Koppula, R. Gupta y A. Saxena, “Aprendiendo las actividades humanas y las posibilidades de los objetos a partir de videos rgb-d”, Int. Revista de investigación en robótica, vol. 32, núm. 8, págs. 951–970, 2013.

[8] Y. Iwai, K. Watanabe, Y. Yagi y M. Yachida, “Reconocimiento de gestos mediante el uso de guantes de colores”, en IEEE Int. Conf. sobre sistemas, hombre y cibernética, vol. 1, 1996, págs. 76–81.

[9] L. Bretzner, I. Laptev y T. Lindeberg, “Reconocimiento de gestos manuales utilizando características de color de múltiples escalas, modelos jerárquicos y filtrado de partículas”, en IEEE Int. Conf. sobre reconocimiento automático de rostros y gestos, 2002, págs. 423–428.

[10] R. Muñoz-Salinas, R. Medina-Carnicer, FJ Madrid-Cuevas y A. Carmona-Poyato, “Siluetas de profundidad para el reconocimiento de gestos”, Cartas de reconocimiento de patrones, vol. 29, núm. 3, págs. 319–329, 2008.

[11] A. Kuznetsova, L. Leal-Taixe y B. Rosenhahn, “Reconocimiento de lenguaje de señas en tiempo real utilizando una cámara de profundidad para el consumidor”, en IEEE Int. Conf. sobre talleres de visión por computadora, 2013, págs.

[12] N. Pugeault y R. Bowden, “Deletrearlo: reconocimiento de ortografía manual en tiempo real”, en IEEE Int. Conf. sobre talleres de visión por computadora, 2011, págs. 1114-1119.

[13] E. Ohn-Bar y MM Trivedi, “Reconocimiento de gestos manuales en tiempo real para interfaces automotrices: un enfoque y evaluaciones multimodales basados en visión”, IEEE Trans. sobre sistemas de transporte inteligentes, vol. 15, núm. 6, págs. 2368-2377, 2014.

[14] P. Molchanov, S. Gupta, K. Kim y J. Kautz, “Reconocimiento de gestos manuales con redes neuronales convolucionales 3D”, en IEEE Conf. sobre talleres de visión por computadora y reconocimiento de patrones, 2015, págs.

[15] J. Nagi, F. Ducatelle, GA Di Caro, D. Cires¸an, U. Meier, A. Giusti, F. Nagi, J. Schmidhuber y LM Gambardella, “Redes neuronales convolucionales de agrupación máxima para el reconocimiento de gestos con las manos basado en la visión”, en IEEE Int. Conf. sobre aplicaciones de procesamiento de imágenes y señales, 2011, págs. 342–347.

[16] X. Chen, H. Guo, G. Wang y L. Zhang, “Red neuronal recurrente aumentada con función de movimiento para el reconocimiento dinámico de gestos con las manos basado en esqueletos”, Computing Research Repository, vol. abs/1708.03278, 2017. [En línea]. Disponible:-http://arxiv.org/abs/1708.03278 [17] JC Núñez, R. Cabido, JJ Pantrigo, AS Montemayor y JF Vélez, “Redes neuronales convolucionales y memoria a largo plazo para la actividad humana basada en esqueletos y el reconocimiento de gestos con las manos”, Pattern Recognition, vol. 76, págs. 80 – 94, 2018.

[18] Q. De Smedt, H. Wannous y JP Vandeborre, “Reconocimiento dinámico de gestos de la mano basado en esqueleto”, en IEEE Conf. sobre talleres de visión por computadora y reconocimiento de patrones, junio de 2016, págs.

[19] MD Zeiler, “Adadelta: un método de tasa de aprendizaje adaptativo”, preimpresión de arXiv arXiv:1212.5701, 2012.

[20] D. Kingma y J. Ba, "Adam: un método para la optimización estocástica", Preimpresión de arXiv arXiv:1412.6980, 2014.