

APRENDIZAJE NO CONTRASTIVO Y NO SUPERVISADO DE SEÑALES FISIOLÓGICAS A PARTIR DE VIDEO

INFORMACIÓN PRIORITARIA

Esta solicitud reivindica el beneficio de la solicitud de patente provisional de Estados Unidos No. 63/424,606 presentada el 11 de noviembre de 2022, que se incorpora en este documento por referencia en su totalidad.

CAMPO DE LA INVENCION

Las modalidades de la presente invención generalmente se relacionan con el uso de biometría, y más particularmente, con el aprendizaje no contrastivo y no supervisado de señales fisiológicas del video.

DISCUSIÓN DE LA TÉCNICA RELACIONADA

En general, la biometría puede usarse para rastrear signos vitales que proporcionan indicadores sobre el estado físico de un sujeto que puede usarse de varias maneras. Como ejemplo, para la seguridad fronteriza o el monitoreo de la salud, se pueden usar signos vitales para detectar riesgos para la salud (por ejemplo, temperatura) o detectar el engaño (por ejemplo, cambio en el pulso o el diámetro de la pupila). Si bien la detección de la temperatura es una tecnología bien desarrollada, la recolección de otros signos vitales útiles y precisos, como la frecuencia del pulso (es decir, la frecuencia cardíaca o los latidos cardíacos por minuto) o la forma de onda de pulso han requerido que los dispositivos físicos se conecten al sujeto. El deseo de realizar medidas biométricas sin contacto físico ha producido algunas técnicas basadas en video.

El desempeño de la estimación de la forma de onda de pulso o la velocidad de pulso confiable desde un sensor de cámara es más difícil que la pletismografía de contacto por varias razones. El cambio en la luz reflejada de la superficie de la piel, debido a la absorción de la luz de la sangre, es muy menor en

comparación con los causados por los cambios en la iluminación. Incluso en configuraciones con iluminación ambiental, los movimientos del sujeto cambian drásticamente la luz reflejada y domina la señal de pulso.

La estimación de los signos vitales a base de cámara es un campo de rápido crecimiento que permite el monitoreo de la salud sin contacto en una variedad de configuraciones. Si bien el número de enfoques exitosos ha aumentado rápidamente, el tamaño de los conjuntos de datos de video de referencia con grabaciones simultáneas de los signos vitales se ha mantenido relativamente estancada. Es bien conocido en toda la comunidad de aprendizaje automático que aumentar la cantidad y la diversidad de los datos de capacitación es una estrategia efectiva para mejorar el rendimiento.

La recopilación de datos fisiológicos remotos es un desafío por varias razones. Primero, la grabación de muchas horas de videos de alta calidad da como resultado un volumen de datos difíciles de manejar. En segundo lugar, registrar una población diversa de sujetos con datos médicos asociados es difícil debido a las preocupaciones de privacidad. Además, las mediciones de contacto sincronizantes con grabaciones de video en diversos entornos dependen en gran medida de la infraestructura de hardware y la configuración de laboratorio del investigador. Incluso las mediciones de contacto utilizadas como referencia contienen ruido, lo que dificulta la curación de datos. Estas dificultades que contribuyen a la depuración de datos, limitan la escalabilidad y la robustez de los modelos.

Hay pocas fuentes disponibles públicamente de videos simultáneos y grabaciones fisiológicas, lo que dificulta la extracción de datos para la capacitación supervisada. Afortunadamente, los trabajos recientes han demostrado que la capacitación no supervisada para fotopletiografía remota (rPPG) es efectiva.

La clase principal de enfoques para la estimación remota del pulso ha cambiado en la última década de la separación de la fuente ciega, a través de

transformaciones de color lineal hasta modelos supervisados basados en el aprendizaje profundo. Si bien las transformaciones de color se generalizan bien en muchos conjuntos de datos, los modelos basados en el aprendizaje profundo dan una mejor precisión cuando se prueban en datos de una distribución similar al conjunto de capacitación. Con este fin, la investigación en aprendizaje profundo se ha centrado en optimizar las arquitecturas neuronales para extraer características espaciales y temporales robustas de los conjuntos de datos de referencia limitados.

Para superar el cuello de botella de datos, los grandes conjuntos de datos fisiológicos sintéticos se han vuelto recientemente más populares. El conjunto de datos SCAMPS contiene videos para 2,800 avatares sintéticos en varios entornos con un rango de etiquetas correspondientes que incluyen PPG, EKG, respiración y unidades de acción facial. El conjunto de datos sintético UCLA contiene 480 videos, y muestran que los modelos de capacitación con datos reales y sintéticos ofrecen los mejores resultados. Otra fuerza de los conjuntos de datos sintéticos es su capacidad para cubrir el rango amplio de tonos de piel, lo que puede ser difícil al recopilar datos reales.

Otra posible solución potencial a la falta de datos de capacitación fisiológica es el aprendizaje no supervisado, donde es suficiente un gran conjunto de videos y antecedentes periódicos en la señal de salida.

El aprendizaje auto-supervisado está progresando rápidamente para el aprendizaje de representación de imágenes. Recientemente, dos clases principales de enfoques han estado compitiendo: el aprendizaje contrastivo y no contrastivo (o regularizado). Los enfoques contrastivos definen los criterios para distinguir si dos muestras son iguales o diferentes, luego comparan las representaciones para extraer o separar las predicciones. Los métodos no contrastivos aumentan los pares positivos y hacen cumplir la varianza en las predicciones sobre los lotes para evitar el colapso, en el que las representaciones del modelo residen en un pequeño subespacio del espacio de características, en lugar de abarcar un espacio de representación más grande

o completo. Los métodos de destilación solo usan muestras positivas y evitan el colapso aplicando una operación de promedio móvil y un operador de gradiente de parada. Otra clase de enfoques maximiza el contenido de información de las representaciones.

Todos los enfoques rPPG no supervisados existentes son contrastivos. En el marco contrastivo, los pares de videos de entrada se pasan como entrada al mismo modelo, y las predicciones en videos similares se acercan, mientras que las predicciones de videos diferentes se repelen. El método para seleccionar muestras negativas difiere entre la literatura anterior, pero el marco contrastivo subyacente es similar.

En consecuencia, los inventores han desarrollado sistemas, dispositivos, métodos e instrucciones no transitorias legibles por computadora que permiten un aprendizaje no contrastivo y no supervisado de las señales fisiológicas del video.

RESUMEN DE LA INVENCION

En consecuencia, la presente invención se dirige al aprendizaje no contrastivo y no supervisado de las señales fisiológicas del video que obvia sustancialmente uno o más problemas debido a limitaciones y desventajas de la técnica relacionada.

Las características y ventajas adicionales de la invención se establecerán en la descripción que sigue, y en parte será evidente a partir de la descripción, o se puede aprender mediante la práctica de la invención. Los objetivos y otras ventajas de la invención serán realizados y alcanzados mediante la estructura particularmente señalados en la descripción escrita y las reivindicaciones del presente, así como los dibujos adjuntos.

Para lograr estas y otras ventajas y de acuerdo con el propósito de la presente invención, como se realizan y se describen en general, las modalidades

incluyen sistemas, dispositivos, métodos e instrucciones no transitorias legibles por computadora para el aprendizaje no contrastivo y no supervisado de una señal fisiológica de una transmisión de video, incluyendo el método implementado por computadora capturar la transmisión de video de un sujeto, incluyendo la transmisión de video una secuencia de cuadros; procesar cada cuadro de la transmisión de video para actualizar una función de detección de señales fisiológicas; determinar la señal fisiológica de la transmisión de video y aplicar la función de detección de señal fisiológica actualizada a una transmisión de video posterior.

En relación con cualquiera de las diversas modalidades, la transmisión de medios o video incluye una o más de una transmisión de video de luz visible, una transmisión de video de infrarrojo cercano, una transmisión de video de infrarrojo de onda larga, una transmisión de video térmico y una transmisión de audio del sujeto.

En relación con cualquiera de las diversas modalidades, la señal fisiológica incluye por lo menos una de la frecuencia de pulso, la presión arterial o la velocidad de parpadeo de los ojos.

En relación con cualquiera de las diversas modalidades, la señal fisiológica incluye por lo menos uno de frecuencia de pulso o frecuencia de voz.

En relación con cualquiera de las diversas modalidades, recortando cada cuadro de la corriente de medios para encapsular una región de interés que incluye una o más de una cara, mejilla, frente u ojo.

En relación con cualquiera de las diversas modalidades, la región de interés incluye dos o más partes del cuerpo.

En relación con cualquiera de las diversas modalidades, combinando por lo menos dos de una transmisión de video de luz visible, una transmisión de video

de infrarrojo cercano y una transmisión de video térmico en una transmisión de video fusionado.

En relación con cualquiera de las diversas modalidades, la transmisión de video de luz visible, la transmisión de video de infrarrojo cercano y/o la transmisión de video térmico se combinan de acuerdo con un dispositivo de sincronización.

Se debe entender que tanto la descripción general anterior y la siguiente descripción detallada son ejemplos y explicativos y están destinados a proporcionar una explicación adicional de la invención como se reivindica.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

Los dibujos que se acompañan, que se incluyen para proporcionar una comprensión adicional de la invención y se incorporan y constituyen una parte de esta especificación, ilustran las modalidades de la invención y, junto con la descripción, sirven para explicar los principios de la invención.

La Figura 1 ilustra un sistema para la estimación de la forma de onda de pulso.

La Figura 2 es una descripción general del marco de aprendizaje no contrastivo y no supervisado (NCUL) para fotoplethysmografía remota (rPPG) en comparación con el aprendizaje tradicional supervisado y no supervisado.

La Figura 3 ilustra el resultado de la capacitación exclusivamente con la pérdida de ancho de banda L_b .

La Figura 4 ilustra los resultados de los modelos capacitados y probados en particiones disjuntas entre sujetos de los mismos conjuntos de datos.

La Figura 5 ilustra las predicciones de forma de onda dentro del conjuntos de datos en conjuntos de datos de referencia de modelos no supervisados de extremo a extremo en una ventana de 8 segundos.

La Figura 6 ilustra los resultados para NCUL y la capacitación supervisada sobre la misma arquitectura.

La Figura 7 ilustra los resultados para la capacitación y las pruebas en UBFC-rPPG.

La Figura 8 ilustra un método implementado por computadora para el aprendizaje no contrastivo y no supervisado de una señal fisiológica de una transmisión de video.

DESCRIPCIÓN DETALLADA DE LAS MODALIDADES EJEMPLARES

La referencia ahora se hará en detalle a las modalidades de la presente invención, cuyos ejemplos se ilustran en los dibujos que se acompañan. Siempre que sea posible, números de referencia similares se utilizarán para elementos similares.

Se describen modalidades de interfaces de usuario y métodos asociados para el uso de un dispositivo. Sin embargo, debe entenderse que las interfaces de usuario y los métodos asociados pueden aplicarse a numerosos tipos de dispositivos, como dispositivos de comunicación portátiles como tabletas o teléfonos móviles. El dispositivo de comunicación portátil admite diversas aplicaciones, como comunicaciones por cable o inalámbricas. Las diversas aplicaciones que pueden almacenarse (en una memoria no transitoria) y ejecutarse (por un procesador) en el dispositivo pueden utilizar por lo menos un dispositivo físico de interfaz de usuario común, como una pantalla táctil. Una o más funciones de la pantalla táctil, así como la información correspondiente que se muestra en el dispositivo, pueden ajustarse y/o modificarse de una aplicación a otra y/o dentro de la aplicación correspondiente. De esta manera,

una arquitectura física común del dispositivo puede admitir diversas aplicaciones con interfaces de usuario que son intuitivas y transparentes.

Las modalidades de la presente invención proporcionan sistemas, dispositivos, métodos e instrucciones no transitorias legibles por computadora para medir uno o más datos biométricos, incluyendo la frecuencia cardíaca, la forma de onda del pulso y/o la respiración, sin contacto físico con el sujeto. Otros datos biométricos pueden incluir el pulso, la mirada, el parpadeo, la pupilometría, la temperatura facial, el nivel de oxígeno, la presión arterial, el audio, el tono y/o la frecuencia de la voz, las microexpresiones, etc. En las diversas modalidades, los sistemas, dispositivos, métodos e instrucciones recopilan, procesan y analizan el video tomado en una o más modalidades (por ejemplo, luz visible, infrarrojo cercano, infrarrojo de onda larga, térmico) para proporcionar un aprendizaje no contrastivo y no supervisado de señales fisiológicas a partir de una señal de video o datos de video (por ejemplo, MP4).

Se pueden utilizar sensores biométricos adicionales para ampliar el potencial y abordar los desafíos de la monitorización remota de personas. En diversas modalidades, también se pueden utilizar cambios en la mirada del sujeto, la frecuencia de parpadeo, el diámetro de la pupila, el habla, la temperatura facial y las microexpresiones.

Como se describe en este documento, el pulso o la forma de onda del pulso del latido cardíaco del sujeto puede utilizarse como dato biométrico para establecer características del estado físico del sujeto y cómo cambian durante un período de observación (por ejemplo, durante un interrogatorio u otra actividad). La fotoplethismografía remota (rPPG) consiste en la monitorización del pulso del volumen sanguíneo desde una cámara a distancia. Mediante la rPPG, se puede detectar el pulso del volumen sanguíneo a partir de un video a una distancia de la superficie de la piel. La divulgación de la Solicitud Estadounidense No. 17/591,929, titulada "VIDEO BASED DETECTION OF PULSE WAVEFORM" (DETECCIÓN DE LA FORMA DE ONDA DEL PULSO

BASADA EN VIDEO), presentada el 3 de febrero de 2022, se incorpora por referencia en su totalidad.

La Figura 1 ilustra un sistema 100 para la estimación de la forma de onda del pulso. El sistema 100 incluye un sistema de sensor óptico 1, un sistema de I/O (Entrada/Salida) de video 6 y un sistema de procesamiento de video 101.

El sistema de sensor óptico 1 incluye uno o más sensores de cámara, cada sensor de cámara respectivo configurado para capturar una transmisión de video que incluye una secuencia de cuadros. Por ejemplo, el sistema de sensor óptico 1 puede incluir una cámara de luz visible 2, una cámara de infrarrojo cercano 3, una cámara térmica 4 o cualquier combinación de estas. En caso de utilizar varios sensores de cámara (por ejemplo, modalidad simple o modalidad múltiple), las múltiples transmisiones de video resultantes pueden sincronizarse de acuerdo con el dispositivo de sincronización 5. Como alternativa, o adicionalmente, se pueden utilizar una o más técnicas de análisis de video para sincronizar las transmisiones de video. Si bien se incluyen una cámara de luz visible 2, una cámara de infrarrojo cercano 3 y una cámara térmica 4, se pueden utilizar otros dispositivos multimedia, como una grabadora de voz.

El sistema de I/O de video 6 recibe una o más transmisiones de video capturadas. Por ejemplo, el sistema I/O de video 6 está configurado para recibir la transmisión de video de luz visible 7, la transmisión de video de infrarrojo cercano 8 y la transmisión de video térmico 9 del sistema de sensores ópticos 1. En este documento, las transmisiones de video recibidas pueden almacenarse de acuerdo con formato(s) digital(es) conocido(s). En caso de recibir varias transmisiones de video (por ejemplo, modalidad simple o modalidad múltiple), el procesador de fusión 10 está configurado para combinar las transmisiones de video recibidas. Por ejemplo, el procesador de fusión 10 puede combinar la transmisión de video de luz visible 7, la transmisión de video de infrarrojo cercano 8 y/o la transmisión de video térmico 9 en una transferencia de video fusionado 11. En este documento, las transmisiones

respectivas pueden sincronizarse según la salida (por ejemplo, una señal de reloj) del dispositivo de sincronización 5.

En el sistema de procesamiento de video 101, el detector de región de interés 12 detecta (es decir, localiza espacialmente) una o más regiones espaciales de interés (ROI) dentro de cada cuadro de video. La ROI puede ser una cara, otra parte del cuerpo (por ejemplo, una mano, un brazo, un pie, un cuello, etc.) o cualquier combinación de partes del cuerpo. Inicialmente, el detector de región de interés 12 determina una o más ROIs espaciales aproximadas dentro de cada cuadro de video. El detector de región de interés 12 es robusto a oclusiones faciales fuertes de mascarillas faciales y otras prendas para la cabeza. Posteriormente, el preprocesador de cuadros 13 recorta el cuadro para encapsular la o las ROI. En algunas implementaciones, el recorte incluye la reducción del tamaño de cada cuadro mediante interpolación bicúbica para reducir el número de píxeles de la imagen a procesar. Alternativamente, o adicionalmente, el cuadro recortado puede redimensionarse aún más para obtener una imagen más pequeña.

El sistema de preparación de secuencias 14 agrega lotes de secuencias o subsecuencias de cuadros ordenados desde el procesador de cuadros 13 para su procesamiento. A continuación, la Red Neuronal Convolutiva Tridimensional (3DCNN) 15 recibe la secuencia o subsecuencia de cuadros del sistema de preparación de secuencias 14. La 3DCNN 15 procesa la secuencia o subsecuencia de cuadros mediante una red neuronal convolutiva tridimensional para determinar las dimensiones espaciales y temporales de cada cuadro y generar un punto de forma de onda de pulso para cada cuadro de la secuencia de cuadros. La 3DCNN 15 aplica una serie de convolución tridimensional, promediado, agrupamiento y no linealidades para producir una señal unidimensional que se aproxima a la forma de onda del pulso 16 de la secuencia o subsecuencias de entrada.

En algunas configuraciones, el sistema de agregación de pulsos 17 combina cualquier número de formas de onda del pulso 16 de las secuencias o

subsecuencias de cuadros en una forma de onda del pulso agregada 18 para representar la transmisión de video completa. El extractor de diagnóstico 19 está configurado para calcular la frecuencia cardíaca y la variabilidad de la frecuencia cardíaca a partir de la forma de onda del pulso agregada 18. Para identificar la variabilidad de la frecuencia cardíaca, se puede comparar la frecuencia cardíaca calculada de varias subsecuencias. La unidad de visualización 20 recibe actualizaciones en tiempo real o casi real del extractor de diagnóstico 19 y muestra la forma de onda del pulso agregada 18, la frecuencia cardíaca y la variabilidad de la frecuencia cardíaca a un operador. La unidad de almacenamiento 21 está configurada para almacenar la forma de onda del pulso agregada 18, la frecuencia cardíaca y la variabilidad de la frecuencia cardíaca asociadas al sujeto.

Adicionalmente, o alternativamente, la secuencia de cuadros puede dividirse en subsecuencias parcialmente superpuestas dentro del sistema de preparación de secuencias 14, donde una primera subsecuencia de cuadros se superpone con una segunda subsecuencia de cuadros. La superposición de cuadros entre subsecuencias evita los efectos de borde. En este caso, el sistema de agregación de pulsos 17 puede aplicar una función de Hann a cada subsecuencia y sumar las subsecuencias superpuestas para generar una forma de onda de pulso agregada 18 con el mismo número de muestras que los cuadros de la transmisión de video original. En algunas configuraciones, cada subsecuencia se pasa individualmente a la 3DCNN 15, que realiza una serie de operaciones para generar una forma de onda de pulso para cada subsecuencia 16. Cada forma de onda de pulso emitida por el 3DCNN 15 es una serie temporal con un valor real para cada cuadro de video. Dado que la 3DCNN 15 procesa cada subsecuencia individualmente, se recombinan posteriormente.

En algunas implementaciones, se pueden aplicar uno o más filtros a la región de interés. Por ejemplo, se pueden filtrar una o más longitudes de onda de luz LED. El LED puede proyectarse sobre toda la región de interés y las superficies circundantes o partes de esta. Adicionalmente, o alternativamente, se pueden

procesar aún más las señales temporales en regiones no cutáneas. Por ejemplo, el análisis de las cejas o la esclerótica del ojo puede identificar cambios fuertemente correlacionados con el movimiento, pero no necesariamente correlacionados con el fotopletiśmograma. Si la misma señal periódica predicha como pulso se encuentra en superficies no cutáneas, puede indicar un sujeto no real o un intento de violación de la seguridad.

Aunque se ilustra como un sistema único, la funcionalidad del sistema 100 puede implementarse como un sistema distribuido. Mientras que el sistema 100 determina la frecuencia cardíaca, otras configuraciones distribuidas rastrean los cambios en la mirada del sujeto, la frecuencia de parpadeo, el diámetro de la pupila, el habla, la temperatura facial y las microexpresiones, por ejemplo. Además, la funcionalidad divulgada en el presente documento puede implementarse en servidores o dispositivos separados que pueden estar conectados entre sí a través de una red, como un quiosco de seguridad conectado a un servidor de backend. Además, es posible que uno o más componentes del sistema 100 no estén incluidos. Por ejemplo, el sistema 100 puede ser un teléfono inteligente o una tableta que incluye un procesador, memoria y una pantalla, pero puede no incluir uno o más de los otros componentes mostrados en la Figura 1. Las modalidades pueden implementarse utilizando diversos dispositivos de procesamiento y almacenamiento de memoria. Por ejemplo, se puede utilizar una CPU y/o un GPU en el sistema de procesamiento para reducir el tiempo de ejecución y calcular el pulso casi en tiempo real. El sistema 100 puede formar parte de un sistema mayor. Por lo tanto, el sistema 100 puede incluir uno o más módulos funcionales adicionales.

Se pueden extraer señales fisiológicas cuasiperiódicas sutiles, como el pulso de volumen sanguíneo y la respiración, del video RGB, lo que permite la monitorización remota de la salud y otras aplicaciones. Los avances en la estimación remota del pulso, o fotopletiśmografía remota (rPPG), se basan actualmente en soluciones de aprendizaje profundo supervisado. Sin embargo, los enfoques actuales se entrenan y evalúan con conjuntos de datos de

referencia limitados, registrados con datos de referencia reales de sensores de contacto PPG.

Las implementaciones proporcionan el primer marco de aprendizaje no contrastivo y no supervisado para la regresión de señales, con el fin de reducir y/o eliminar las limitaciones de los datos de video etiquetados. Con suposiciones mínimas de periodicidad y ancho de banda finito, las modalidades identifican un pulso de volumen sanguíneo directamente a partir de videos no etiquetados. Fomentar espectros de potencia dispersos dentro de los límites de banda deseados y la varianza en un lote de espectros de potencia son suficientes para el aprendizaje de las características visuales de las señales periódicas. Se validó el uso de datos de video no etiquetados, no creados específicamente para rPPG, para entrenar estimadores robustos de la frecuencia del pulso. Dados los limitados sesgos inductivos y los resultados empíricos positivos, las modalidades se pueden aplicar fácilmente a otras señales periódicas de video, lo que permite múltiples mediciones fisiológicas sin necesidad de datos de referencia reales.

Las modalidades proporcionan aprendizaje no contrastivo y no supervisado de señales fisiológicas a partir de un video. Un modelo puede ser guiado hacia la extracción de señales periódicas del video utilizando una formulación no contrastiva. Al proporcionar videos faciales y recibir instrucciones para identificar una señal periódica entre 40 y 180 latidos por minuto, el modelo aprende correctamente a estimar el pulso de volumen sanguíneo, a pesar de la sutileza de las características faciales visuales que proporcionan información sobre el pulso sanguíneo.

Las modalidades proporcionan un marco para el aprendizaje no contrastivo y no supervisado mediante el aprovechamiento de las señales periódicas previas. Las modalidades también proporcionan el primer método de aprendizaje no contrastivo y no supervisado para la medición de signos vitales con cámara. Además, las modalidades permiten capacitar modelos con un

conjunto de datos de video no específico para rPPG sin signos vitales de valores reales.

La Figura 2 muestra una descripción general del marco de aprendizaje no contrastivo y no supervisado (NCUL) para la fotopletiografía remota (rPPG) en comparación con el aprendizaje tradicional supervisado y no supervisado. Las pérdidas supervisadas y contrastivas utilizan métricas de distancia a los datos de referencia u otras muestras. El marco aplica la pérdida directamente a la predicción mediante la configuración del espectro de frecuencias y el fomento de la varianza en un lote de entradas. La potencia fuera de los límites de banda se penaliza para aprender invariancias a frecuencias irrelevantes. Se fomenta que la potencia dentro de los límites de banda se distribuya dispersamente cerca de la frecuencia máxima.

Inicialmente, se formula la configuración general para la regresión de señales a partir de un video. Una muestra de video $x_i \in \mathbb{R}^{T \times W \times H \times C}$, muestreada a partir de un conjunto de datos D , consta de T imágenes de tamaño $W \times H$ píxeles en C canales, capturadas uniformemente a lo largo del tiempo. Los métodos del estado de la técnica ofrecen modelos f que estiman mediante regresión una forma de onda $\mathbb{R}^T \ni y_i = f(x_i)$ de la misma longitud que el video. Recientemente, la tarea se ha modelado eficazmente de extremo a extremo con los modelos f como redes neuronales espaciotemporales. Mientras que trabajos anteriores son supervisados y minimizan la pérdida en la medición de un pulso de contacto, en este caso el aprendizaje no contrastivo utiliza únicamente la estimación del modelo.

Se pueden establecer valores a priori sólidos en el pulso estimado (por ejemplo, rango de frecuencia y alta periodicidad). A continuación, se pueden implementar una o más restricciones en el dominio de la frecuencia en lugar del dominio del tiempo. De este modo, las predicciones de la forma de onda se pasan por la FFT antes de calcular las pérdidas.

Pérdidas. Una de las ventajas del aprendizaje no supervisado para señales periódicas es que el espacio de soluciones se puede restringir

significativamente. Para señales fisiológicas como la respiración y el pulso del volumen sanguíneo, se conocen los límites superior e inferior saludables de las frecuencias en respiraciones y latidos por minuto. La señal extraída es relativamente dispersa en el dominio de la frecuencia, y el modelo filtra las señales de ruido presentes en el video. Con estas restricciones, se simplifica el problema de encontrar características adecuadas para la señal deseada en los datos.

Una de las restricciones más útiles del modelo son los límites de banda de frecuencia. Los métodos no supervisados anteriores han utilizado la razón de potencia irrelevante (IPR) como métrica de validación para la selección de modelos. El uso de límites de banda de frecuencia también es eficaz durante el entrenamiento del modelo. Con límites de banda inferior y superior de a y b , respectivamente, la pérdida de ancho de banda se convierte en:

$$L_b = \frac{\sum_{i=-\infty}^a F_i + \sum_{i=b}^{\infty} F_i}{\sum_{i=-\infty}^{\infty} F_i}, \quad (1)$$

donde F_i es la potencia espectral de la señal predicha para la frecuencia i , lo que permite penalizar el modelo para generar señales fuera de los límites de banda deseados. Esta pérdida hace cumplir el aprendizaje de muchos invariantes, como el movimiento de la respiración, la conversación o las expresiones faciales que generalmente ocupan frecuencias bajas. Por ejemplo, se pueden usar los límites como $a = 0.66$ Hz a $B = 3$ Hz, lo que corresponde a un rango de frecuencia de pulso común de 40 bpm a 180 bpm.

La Figura 3 ilustra el resultado de la capacitación exclusivamente con la pérdida de ancho de banda L_b . Cada columna muestra predicciones de modelos capacitados con una o cada una de las pérdidas para 20 épocas en UBFC-rPPG. Las dos primeras filas ilustran una muestra en el dominio de tiempo y frecuencia, respectivamente. La última fila ilustra la distribución de la potencia de señal en todo el conjunto de validación. La pérdida de ancho de banda penaliza la potencia de señal fuera de los límites de banda predefinidos (por ejemplo, 40 a 180 bpm) para restringir el espacio de salida. La última fila

también ilustra que el modelo limita la potencia de la señal entre los límites de banda. La pérdida de dispersión fomenta un espectro estrecho que contiene un componente periódico fuerte. Por sí mismo, el modelo aprende la solución caracterizada por frecuencias muy bajas. La pérdida de varianza fomenta diversos espectros de potencia en un lote, evitando que el modelo se derrumbe a un ancho de banda estrecho. Cuando se combina, el modelo aprende a estimar las señales periódicas dentro de los límites de banda deseados.

El pulso del volumen sanguíneo contiene una frecuencia dominante, y es el marcador fisiológico más común utilizado en la práctica es la frecuencia del pulso. El modelo se puede mejorar penalizando las predicciones de banda ancha. Esto también simplifica la verdadera señal de que el modelo debe identificarse penalizando la dinámica visual que puede no exudar periodicidad. Se puede usar una formulación similar al IPR, pero los límites de frecuencia se seleccionan cerca de la frecuencia predicha máxima. Concretamente, la pérdida de dispersión, L_s , es la misma que la ecuación (1) si los límites de banda se sustituyen por $-\infty$ y ∞ , $a = \operatorname{argmax}(F) - \Delta F$ y $b = \operatorname{argmax}(F) + \Delta F$. Los experimentos se realizaron con un ΔF de 6 latidos por minuto. Aquí, la Figura 3 ilustra el resultado de capacitar solo con la pérdida de dispersión en la segunda columna. Para una sola muestra, el espectro de potencia es muy disperso. Para todo el conjunto de datos, es más fácil predecir soluciones dispersas cuando las altas frecuencias se ignoran por completo y se aprenden características visuales correspondientes a bajas frecuencias.

Uno de los riesgos de los métodos no contrastivos es el modelo que colapsa en soluciones triviales y hacer predicciones independientemente de las características de entrada. En métodos regularizados como VICReg, se utiliza una pérdida de bisagra en la varianza sobre un lote de predicciones para hacer cumplir diversos resultados. Se puede utilizar una estrategia similar para evitar el colapso del modelo, pero en su lugar difundir la varianza en las densidades espectrales de potencia hacia una distribución uniforme sobre las frecuencias compatibles.

La pérdida de varianza procesa una distribución previa uniforme P sobre las frecuencias d , y un lote de n densidades espectrales, $F = [v_1, \dots, v_n]$, donde cada vector es una descomposición de frecuencia D -dimensional de una forma de onda predicha. Se calcula la suma normalizada de densidades sobre el lote, Q . Además, se define la pérdida de varianza como la distancia cuadrada de Wasserstein al previo uniforme:

$$L_v = \frac{1}{d} \sum_{i=1}^d (\text{CDF}_i(Q) - \text{CDF}_i(P))^2, \quad (2)$$

donde CDF es una función de distribución acumulativa. La pérdida de varianza se beneficia de grandes tamaños de lotes, ya que los lotes pequeños pueden ocupar un pequeño rango de frecuencias. La tercera columna de la Figura 3 ilustra el efecto de la pérdida de varianza durante la capacitación del modelo. Para una sola muestra, se predicen señales de banda ancha que contienen múltiples frecuencias, y las frecuencias sobre el conjunto de datos predicho cubren el ancho de banda de la tarea.

Resumiendo, la función de pérdida de capacitación es una suma de las pérdidas antes mencionadas:

$$L = L_b + L_s + L_v. \quad (3)$$

Aunque se podría ponderar ciertos componentes de la pérdida más que otros, las pérdidas fueron formuladas para escalar entre 0 y 1. Según el uso experimental, se descubrió que la suma sin ponderación da un buen rendimiento. La función de pérdida combinada alienta al modelo a buscar sobre las frecuencias compatibles para descubrir características visuales para una señal periódica fuerte. Sorprendentemente, el marco es suficiente para aprender a retroceder el volumen de sangre en el video, como se ilustra en la última columna de la Figura 3.

A diferencia de los enfoques conocidos, que solo aplican aumentos de frecuencia, se aplican varios aumentos tanto a las dimensiones espaciales como a las dimensiones temporales para mejorar el modelo para identificar las invariancias al ruido que puede encontrar en configuraciones reales.

Aumentos de intensidad de imagen. Se agrega ruido gaussiano aleatorio a cada ubicación de píxeles en un clip con una media de 0 y una desviación estándar de 2 en la escala de imagen original de 0 a 255. La iluminación se aumenta al agregar una muestra constante de una distribución gaussiana con media 0 y desviación estándar de 10 a cada píxel en un clip, que oscurece o aclara el video.

Aumentos espaciales. Un videoclip se recorta de forma horizontal en forma cuadrada con una probabilidad del 50%. La dimensión espacial de un clip se recorta aleatoriamente en forma cuadrada a un tamaño comprendido entre la mitad de la longitud original y la longitud original. El clip recortado se interpola linealmente hasta las dimensiones espaciales originales.

Aumentos temporales. Con la suposición general de que la señal deseada es fuertemente periódica y escasamente representada en el dominio de Fourier, un videoclip se voltea aleatoriamente a lo largo de la dimensión de tiempo con una probabilidad del 50%. Tenga en cuenta que la descomposición de Fourier de una senoide con inversión de tiempo es idéntica a la de la senoide original.

Aumentos de frecuencia. Quizás el aumento más importante es el remuestreo de frecuencia, donde el video se interpola linealmente a una velocidad de cuadro diferente. Este aumento es particularmente interesante para rPPG, ya que transforma la entrada de video y la señal de destino de manera equivalente a lo largo de la dimensión del tiempo, lo que lo hace equivariante. Dadas las transformaciones antes mencionadas que son invariantes, $\tau(\cdot) \sim T$, la operación de remuestreo de frecuencia equivariante, $\phi(\cdot) \Phi$ y un modelo $f(\cdot)$ que infiere una forma de onda de un video que tiene lo siguiente:

$$\phi(f(\tau(x))) = f(\phi(\tau(x))). \quad (4)$$

Este es un aumento poderoso, porque permite el aumento de la distribución de destino junto con la entrada de video. En los experimentos, los clips de entrada se volvieron a muestrear aleatoriamente mediante un factor $c \sim U(0.6, 1.4)$.

Después de aplicar el aumento de remuestreo, los límites de banda se escalaron posteriormente mediante c para evitar penalizar el modelo si el aumento empujaba la frecuencia de pulso subyacente fuera de los límites de banda originales.

Se realizaron varios experimentos para evaluar el rendimiento dentro del conjunto de datos y entre diferentes conjuntos de datos. PURE, UBFC-rPPG y DDPM se usaron como conjuntos de datos rPPG de referencia tanto para capacitación como para pruebas, y el conjunto de datos CelebV-HQ y HKBU-MARs solo para capacitación no supervisada.

La Detección del Engaño y Monitoreo Fisiológico (DDPM) consta de 86 sujetos en un entorno de entrevista, donde los sujetos intentaron responder a las preguntas engañosamente. Las entrevistas se registraron a 90 cuadros por segundo durante más de 10 minutos en promedio. La conversación natural y los cambios frecuentes en la pose de la cabeza lo convierten en un conjunto de datos rPPG difícil y menos restringido.

PURE es un conjunto de datos rPPG de referencia que consta de 10 sujetos registrados en 6 sesiones. Cada sesión duró aproximadamente 1 minuto, y el video sin editar se grabó a 30 fps. Las 6 sesiones para cada sujeto consistieron en: (1) estable, (2) conversando, (3) traslado lento de la cabeza, (4) traslado rápido de la cabeza, rotaciones de cabeza pequeña (5) y mediana (6). Las velocidades de pulso están en o cerca de la tasa de descanso del sujeto.

UBFC-rPPG contiene videos de 1 minuto de 42 sujetos grabados a 30 fps. Los sujetos jugaron un juego matemático sensible al tiempo para elevar sus frecuencias cardíacas, pero el movimiento de la cabeza es limitado durante la grabación.

El Ataque con Máscara 3D HKBU con Variaciones del Mundo Real (HKBU-MARs) consta de 12 sujetos capturados en 6 configuraciones de iluminación diferentes con 7 cámaras diferentes cada una, lo que resulta en 504 videos que

duran 10 segundos cada uno. Los diversos sensores de iluminación y cámara lo convierten en un valioso conjunto de datos para la capacitación no supervisada. Se usó la versión 2 de HKBU-MARs, que contiene videos con máscaras 3D realistas y sujetos sin máscara.

El Conjunto de Datos de Video de Celebridades de Alta Calidad (CelebVHQ) es un conjunto de videos procesados de YouTube que contienen 35,666 videos faciales de más de 15,000 identidades. Los videos varían dramáticamente en longitud, condiciones de iluminación, emoción, movimiento, tonos de piel y sensores de cámara. Dado un método suficiente para el aprendizaje no supervisado, el conjunto de datos CelebV-HQ parece un buen candidato para capacitar a un modelo robusto para la estimación de pulso. El mayor desafío para aprovechar los videos en línea es su calidad reducida, que puede comprimirse tanto antes de la carga como por el proveedor de videos. La compresión es un desafío conocido para rPPG, ya que el pulso del volumen de sangre es tan sutil ópticamente.

Capacitación, Preprocesamiento de Datos. Para preparar los videoclips para los modelos de aprendizaje profundo espacio-temporal, se extrajeron 68 puntos de referencia faciales con OpenFace. Se definió una caja delimitadora en cada cuadro con las ubicaciones mínimas y máximas (x, y) extendiendo el cultivo horizontalmente en un 5% para garantizar que las mejillas y la mandíbula estén presentes. La parte superior e inferior se extendió en el 30% y el 5% de la altura de la caja delimitadora, respectivamente, para incluir la frente y la mandíbula. El más corto de los dos ejes se extendió a la longitud del otro para formar un cuadrado. Los cuadros recortados se redimensionaron a 64×64 píxeles con interpolación bicúbica. Para un procesamiento más rápido del conjunto de datos masivo CelebV-HQ, se utilizó MediaPipe Face Mesh para el punto de referencia.

Arquitecturas de Modelos. Se puede utilizar una arquitectura 3D-CNN. Se usó un ancho de núcleo temporal de 5, y el relleno con ceros predeterminado fue reemplazado por la repetición de los bordes. El relleno con ceros a lo largo de

la dimensión de tiempo puede dar lugar a efectos de borde que agregan frecuencias artificiales a las predicciones. Los experimentos mostraron que las dilataciones temporales internas causaban aliasing y reducían el ancho de banda del modelo a frecuencias específicas. Las pérdidas y el marco pueden aplicarse a cualquier tarea y arquitectura con predicciones densas a lo largo de una o más dimensiones. Sin embargo, las arquitecturas populares de rPPG como Deepphys y MTTs-CAN pueden estar mal adecuadas para el enfoque, ya que consumen muy pocos cuadros, y el número de puntos de tiempo debe ser lo suficientemente grande como para dar una resolución de frecuencia suficiente con el FFT.

Capacitación supervisada. Para comparar adecuadamente las modalidades con su contraparte supervisada, se utilizó la misma arquitectura modelo y se capacitó con la pérdida negativa de Pearson entre la forma de onda predicha y la verdad de referencia del sensor de contacto. Durante la capacitación, se aplicaron los mismos aumentos, excepto la reversión del tiempo. Los modelos están entrenados para 200 épocas en PURE y UBFC-rPPG, y para 40 épocas en DDPM. El modelo de la época con la pérdida más baja en el conjunto de validación se selecciona para prueba.

Capacitación no supervisada. Los modelos no supervisados están capacitados para el mismo número de épocas que la configuración supervisada para PURE y UBFCrPPG, pero se capacitan para 40 épocas adicionales en DDPM, ya que este conjunto de datos es considerablemente más difícil. Al contrario de los enfoques no supervisados anteriores, los conjuntos de validación se aprovecharon para la selección del modelo seleccionando el modelo con las pérdidas de paso de banda y de escasez más bajas en muestras invisibles.

Evaluación. Las velocidades de pulso se calculan como el pico espectral más alto entre 0.66 Hz y 3 Hz (equivalente a 40 bpm a 180 lpm) en una ventana deslizante de 10 segundos. El mismo procedimiento se aplica a las formas de onda de referencia para una evaluación confiable. Se aplican métricas de error comunes, como el error absoluto medio (MAE), el error cuadrático medio

(RMSE) y el coeficiente de correlación de Pearson entre las velocidades de pulso. Realizamos una validación cruzada de 5 veces para PURE y UBFC, y usamos las divisiones de conjunto de datos predefinidas de DDPM. Tres (3) modelos fueron con diferentes inicializaciones, lo que dio como resultado 15 modelos capacitados en PURE y UBFC cada uno, y 3 modelos capacitados en DDPM. Se presentan la desviación media y estándar de los errores en los resultados.

Resultados, Prueba de Conjunto de Datos. La Figura 4 ilustra los resultados de los modelos capacitados y probados en particiones disjuntas entre sujetos de los mismos conjuntos de datos. Para PURE y UBFC, se logró un MAE inferior a 1 bpm, que funcionó mejor o a la par con todos los enfoques de aprendizaje tradicionales y supervisados. Para PURE, las modalidades dan un MAE más bajo y un Pearson r de casi 1. El rendimiento cae en DDPM debido a la dificultad general del conjunto de datos y segmentos de ruido del movimiento del oxímetro de los dedos. Sin embargo, las modalidades superan los enfoques no supervisados existentes y los métodos tradicionales, solo que los superan los enfoques supervisados de aprendizaje profundo.

En comparación con otros métodos no supervisados, Contrast-Phys ofrece el rendimiento más competitivo en todos menos DDPM. Tener en cuenta que las modalidades dan el MAE más bajo en todos los conjuntos de datos, pero tiene un RMSE más alto. Es probable que esto se deba a su uso de la eliminación armónica como una etapa posterior al procesamiento al estimar la frecuencia del pulso, que no se describe en, pero se puede encontrar en su código disponible públicamente.

La Figura. 5 ilustra las predicciones de forma de onda dentro del conjunto de datos en conjuntos de datos de referencia de modelos no supervisados de extremo a extremo en una ventana de 8 segundos. Las predicciones del modelo son notablemente periódicas sin ninguna forma de filtrado. Tener en cuenta que la fase no se considera durante la capacitación, por lo que cada modelo aprende su propia fase.

Pruebas entre Conjunto de Datos. Además de los experimentos tradicionales dentro del conjunto de datos, se realizaron pruebas entre conjunto de datos para analizar si el enfoque es robusto a los cambios en la iluminación, el sensor de la cámara, la distribución de la frecuencia de pulso y el movimiento. La Figura 6 ilustra los resultados para NCUL y la capacitación supervisada sobre la misma arquitectura. En general, se descubrió que el rendimiento es similar para los enfoques supervisados y no supervisados cuando se transfieren a diferentes fuentes de datos. La capacitación en PURE exclusivamente da resultados relativamente pobres cuando se transfiere a UBFC-rPPG y DDPM, debido a la baja variabilidad de la frecuencia de pulso y la falta de movimiento. La capacitación en DDPM ofrece los mejores resultados en general, ya que el conjunto de datos es el más grande y captura los movimientos de sujetos más grandes en comparación con otros conjuntos de datos.

Capacitación con videos de Celebv-HQ dada la abundancia de videos faciales disponibles públicamente en línea, un modelo fue capacitado en caras del conjunto de datos Celebv-HQ. Después de procesar los videos disponibles para descargar con MediaPipe y volver a muestrear los videoclips a 30 fps, el conjunto de datos sin etiquetar constaba de 34,029 videos. El modelo capacitó para 23 épocas y fue dejado de capacitar manualmente debido a una meseta en la pérdida de validación. Desafortunadamente, se descubrió que el modelo no podía converger al verdadero pulso del volumen sanguíneo. Es probable que la falla se deba a la mala calidad de los datos de la compresión. Aunque los videos se descargaron con la calidad más alta disponible, los videos probablemente se han comprimido varias veces, eliminando de manera efectiva la señal de pulso por completo. El MAE en bpm para UBFC-rPPG, PURE y DDPM fue 19.22, 24.83 y 27.41, respectivamente.

Capacitación con videos de HKBU-MARs. El conjunto de datos HKBU-MARs fue diseñado para la detección de ataque de presentación de la cara, pero los modelos fueron capacitados en las sesiones de video “reales” en el conjunto de datos. Las filas inferiores en la Figura 6 ilustra los resultados para la capacitación exclusivamente en HKBUMARs, luego se prueba en los conjuntos

de datos rPPG de referencia. La capacitación en HKBU-MARs da mejores resultados al transferir a UBFC-rPPG y PURE que todos los conjuntos de capacitación, excepto DDPM, que es un orden de magnitud más grande y contiene artefactos de movimiento. Este es el primer experimento conocido y exitoso que muestra que los videos no rPPG se pueden usar para capacitar modelos robustos, incluso si no tienen etiquetas de pulso de referencia.

Estudio de Ablación sobre Pérdidas. Los modelos fueron capacitados utilizando todas las combinaciones de componentes de pérdida para analizar sus contribuciones. La Figura 7 ilustra los resultados para el entrenamiento y las pruebas en UBFC-rPPG. La pérdida de paso de banda es crítica para descubrir el pulso del volumen sanguíneo verdadero, mientras que las pérdidas de escasez y varianza no aprenden la señal deseada por sí mismos. Sorprendentemente, combinar la pérdida de paso de banda con solo una de las pérdidas de escasez o varianza da un rendimiento peor que la pérdida de paso de banda. Sin embargo, al combinar los tres componentes, el modelo logra mejores resultados.

Discusión. Inicialmente es sorprendente que la capacitación no supervisada conduzca a modelos de estimación RPPG similares o mejorados en comparación con los capacitados de manera supervisada. Sin embargo, hay varios beneficios potenciales para la capacitación no supervisada. Desde una perspectiva de hardware, una de las dificultades en la capacitación supervisada es alinear la forma de onda de pulso de contacto con los cuadros de video. El sensor de pulso y la cámara pueden tener un retraso de tiempo, dando efectivamente al modelo un objetivo fuera de fase al momento de la capacitación. Un beneficio de la capacitación no supervisada es dar al modelo la libertad de aprender la fase directamente del video.

La señal de pulso de contacto también puede ser ruidosa, ya que PPG es muy sensible al movimiento. Dado que el movimiento puede ocurrir en la cara y la punta del dedo simultáneamente, el ruido del movimiento puede aparecer en el

objetivo, equivocando al modelo para aprender una característica visual para la cual deben ser invariantes.

Desde una perspectiva fisiológica, el pulso observado ópticamente en la punta de los dedos con un sensor de contacto tiene una fase diferente a la de la cara, ya que la sangre se propaga a lo largo de una ruta diferente antes de alcanzar la microvasculatura periférica, lo que hace que la alineación sea casi imposible sin cambiar los objetivos a estimaciones rPPG de los métodos existentes.

Además, la forma morfológica de la forma de onda de contacto PPG depende de numerosos factores, como la longitud de onda de la luz (y la profundidad de penetración de tejidos correspondientes), la presión externa del clip del oxímetro y la vasodilatación en el sitio de medición. Estos factores externos indican que la morfología y la fase de las formas de onda PPG objetivo probablemente sean diferentes de la forma de onda rPPG observada. Al capacitar un modelo para predecir los objetivos proxy, se introdujeron artefactos innecesarios y no se aprendieron características óptimas para una medición precisa de pulso.

El éxito del enfoque no contrastivo propuesto depende de propiedades específicas de los datos, el modelo y cómo interactúan los dos. La capacidad limitada del modelo es en realidad una fortaleza, ya que obliga a descubrir características a generalizarse entre las entradas. Una red de capacidad infinita podría descubrir señales espurias en los datos de capacitación y no generalizar. Al restringir las predicciones del modelo para tener propiedades periódicas específicas, el modelo de capacidad limitada debe encontrar un conjunto general de características para producir una señal que existe en todas las muestras de capacitación, que resulta ser el pulso de volumen sanguíneo en los conjuntos de datos.

Como un efecto secundario beneficioso, el modelo aprende intrínsecamente a ignorar los factores de ruido comunes como la iluminación, el movimiento rígido, el movimiento no rígido (por ejemplo, hablar, sonreír, etc.), y el ruido de

la cámara, ya que pueden dar lugar a señales fuera de los límites de banda predefinidos o con espectros de potencia uniformes. Incluso si el ruido exhibe tendencias periódicas dentro de los límites de la banda para algunas muestras, esas características producirían señales pobres en otras muestras. Por lo tanto, los enfoques no supervisados de extremo a extremo son particularmente adecuados para problemas periódicos.

La Figura 8 ilustra un método implementado por computadora para el aprendizaje no contrastivo y no supervisado de una señal fisiológica de una transmisión de video.

En 810, el método captura una transmisión de medios (por ejemplo, transmisión de video o audio) de un sujeto que incluye la transmisión de medios que incluye una secuencia de cuadros. La transmisión de video puede incluir una o más de una transmisión de video de luz visible, una transmisión de video de infrarrojo cercano y/o una transmisión de video térmico de un sujeto. En algunos casos, el método puede combinar por lo menos dos de la transmisión de video de luz visible, la transmisión de video de infrarrojo cercano y/o la transmisión de video térmico en una transmisión de video fusionado para procesarse. La transmisión de video de luz visible, la transmisión de video de infrarrojo cercano y/o la transmisión de video térmico se combinan de acuerdo con un dispositivo de sincronización y/o una o más técnicas de análisis de video.

A continuación, en 820, el método procesa cada cuadro de la transmisión de medios para actualizar una función de detección de señal fisiológica (por ejemplo, actualiza la función representada como $f(\cdot)$ en la Figura 2, o pesos de la misma).

En 830, el método determina la señal fisiológica de la transmisión de medios. Por ejemplo, la señal fisiológica puede incluir una pluralidad de biometría que incluye frecuencia cardíaca, forma de onda de pulso y/o respiración. Otros ejemplos pueden incluir pulso, mirada, parpadeo, pupilometría, temperatura

facial, nivel de oxígeno, presión arterial, audio, tono de voz y/o frecuencia, microexpresiones, etc.). Aunque no se muestra, la función de detección de señal fisiológica actualizada se puede aplicar fácilmente a una transmisión de video posterior.

Además de las mejoras generales sobre el rendimiento, NCUL también está de extremo a extremo sin supervisión, por lo que no requiere un ajuste posterior de las representaciones aprendidas y, en su lugar, puede inferir las formas de onda directamente. Los aumentos también son mucho más simples de implementar y requieren menos procesamiento durante la capacitación.

En consecuencia, las modalidades introducen un nuevo enfoque de aprendizaje no contrastivo para la regresión de señal no supervisada de extremo a extremo, con experimentos específicos sobre la estimación del pulso del volumen sanguíneo de los videos faciales. Se ha demostrado que una formulación de pérdida simple que requiere restricciones de frecuencia suelta es efectiva para aprender características visuales poderosas.

Será evidente para aquellos hábiles en la técnica que se pueden hacer diversas modificaciones y variaciones en el aprendizaje no contrastivo y no supervisado de señales fisiológicas del video de la presente invención sin apartarse del espíritu o alcance de la invención. Por lo tanto, se pretende que la presente invención cubra las modificaciones y variaciones de esta invención, siempre que se encuentren dentro del alcance de las reivindicaciones adjuntas y sus equivalentes.