Reconocimiento de pose humana en tiempo real en partes a partir de imágenes de profundidad única

Jamie Shotton Andrew Fitzgibbon cocinero estera Toby Sharp Marcos Finocchio
Richard Moore Alex Kipman andres blake
Microsoft Research Cambridge e incubación de Xbox

Abstracto

Proponemos un nuevo método para predecir de forma rápida y precisa las posiciones 3D de las articulaciones del cuerpo a partir de una única imagen de profundidad. sin utilizar información temporal. Adoptamos un enfoque de reconocimiento de objetos, diseñando una representación intermedia de partes del cuerpo que mapea el difícil problema de estimación de pose.
en un problema de clasificación por píxel más simple. Nuestro gran y un conjunto de datos de entrenamiento muy variado permite al clasificador estimar partes del cuerpo invariantes a la pose, forma del cuerpo, vestimenta, etc. Finalmente generamos propuestas 3D con puntuación de confianza de varias articulaciones del cuerpo reproyectando el resultado de la clasificación y encontrar modos locales.

El sistema funciona a 200 fotogramas por segundo en el consumidor. hardware. Nuestra evaluación muestra una alta precisión en ambos conjuntos de pruebas sintéticos y reales, e investiga el efecto de varios parámetros de entrenamiento. Logramos una precisión de última generación en nuestra comparación con trabajos relacionados y demostramos generalización mejorada sobre el esqueleto completo exacto más cercano coincidencia de vecinos.

1. Introducción

El seguimiento interactivo y robusto del cuerpo humano tiene aplicaciones que incluyen juegos, interacción persona-computadora, seguridad, telepresencia e incluso atención médica. Recientemente, la tarea se ha simplificado enormemente con la introducción de cámaras de profundidad en tiempo real [16, 19, 44, 37, 28, 13]. Sin embargo, incluso Los mejores sistemas existentes todavía presentan limitaciones. En particular, hasta el lanzamiento de Kinect [21], ninguno se ejecutaba en modo interactivo. tarifas en hardware de consumo mientras maneja una gama completa de Formas y tamaños del cuerpo humano sometidos a movimientos corporales generales. Algunos sistemas alcanzan altas velocidades siguiendo desde fotograma a fotograma pero tiene dificultades para reinicializarse rápidamente y así no son robustos. En este artículo nos centramos en el reconocimiento de poses. en partes: detectar a partir de una única imagen de profundidad un pequeño conjunto de Candidatos de posición 3D para cada articulación esquelética. Nuestro enfoque en La inicialización y recuperación por cuadro está diseñada para complementar cualquier algoritmo de seguimiento apropiado [7, 39, 16, 42, 13]. que podría incorporar aún más la coherencia temporal y cinemática. El algoritmo presentado aquí forma un componente central de la plataforma de juegos Kinect [21].

Ilustrado en la Fig. 1 e inspirado en un trabajo reciente de reconocimiento de objetos que divide los objetos en partes (por ejemplo, [12, 43]), Nuestro enfoque está impulsado por dos objetivos de diseño clave: eficiencia computacional y robustez. Una única imagen de profundidad de entrada se segmenta en un denso etiquetado probabilístico de partes del cuerpo, con las partes definidas para estar localizadas espacialmente cerca del esqueleto

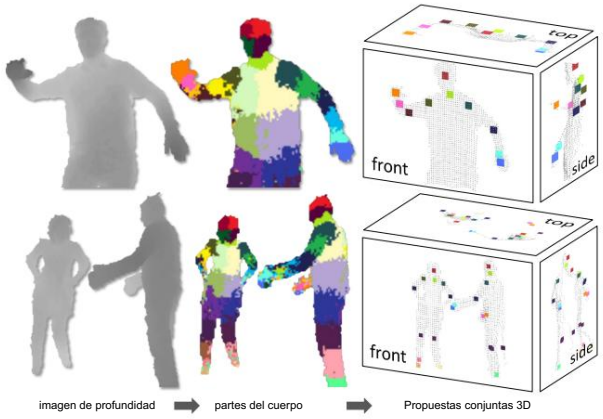

Figura 1. Descripción general. A partir de una única imagen de profundidad de entrada, una imagen por píxel Se infiere la distribución de las partes del cuerpo. (Los colores indican lo más probable etiquetas de piezas en cada píxel, y se corresponden en las propuestas conjuntas). Se estima que los modos locales de esta señal ofrecen propuestas de alta calidad para la ubicación 3D de las articulaciones del cuerpo, incluso para múltiples usuarios.

articulaciones de interés. Reproyectar las partes inferidas en el mundo. espacio, localizamos los modos espaciales de cada distribución de partes. y así generar (posiblemente varios) ponderados por la confianza. propuestas para las ubicaciones 3D de cada articulación esquelética.

Tratamos la segmentación en partes del cuerpo por píxel. tarea de clasificación (ningún término por pares o CRF ha demostrado necesario). Evaluar cada píxel por separado evita una búsqueda combinatoria sobre las diferentes articulaciones del cuerpo, aunque Dentro de una misma parte, por supuesto, todavía existen diferencias dramáticas en la apariencia contextual. Para datos de entrenamiento, generamos imágenes de profundidad sintéticas realistas de humanos de muchas formas y tamaños en poses muy variadas tomadas de una gran base de datos de captura de movimiento. Entrenamos un clasificador de bosque de decisión aleatorio profundo que evita el sobreajuste mediante el uso de cientos de miles de imágenes de entrenamiento. Las características de imagen de comparación de profundidad simples y discriminativas producen Invariancia de traducción 3D manteniendo una alta eficiencia computacional. Para mayor velocidad, se puede ejecutar el clasificador. en paralelo en cada píxel de una GPU [34]. Finalmente, espacial Se calculan los modos de las distribuciones por píxel inferidas. utilizando el desplazamiento medio [10], lo que da como resultado propuestas conjuntas 3D.

Se ejecuta una implementación optimizada de nuestro algoritmo. menos de 5 ms por cuadro (200 cuadros por segundo) en Xbox 360 GPU, al menos un orden de magnitud más rápido que los enfoques existentes. Trabaja cuadro por cuadro en formas y tamaños corporales dramáticamente diferentes, y el enfoque discriminativo aprendido maneja naturalmente las autooclusiones y

Poses recortadas por el marco de la imagen. Evaluamos en ambos reales.
e imágenes de profundidad sintéticas, que contienen poses desafiantes de
un conjunto variado de temas. Incluso sin explotar temporal
o restricciones cinemáticas, las propuestas conjuntas 3D son precisas y estables.
Investigamos el efecto de varios parámetros de entrenamiento y mostramos cómo
los árboles muy profundos aún pueden evitar
sobreajuste debido al gran conjunto de entrenamiento. demostramos
que nuestras propuestas parciales generalicen al menos tan bien como exactamente
vecino más cercano en un entorno tanto idealizado como realista,
y mostrar una mejora sustancial respecto al estado de la
arte. Además, los resultados sobre imágenes de siluetas sugieren una aplicabilidad
más general de nuestro enfoque.

Nuestra principal contribución es tratar la estimación de pose como reconocimiento
de objetos utilizando una novedosa representación de partes intermedias del cuerpo
diseñada para localizar espacialmente las articulaciones de interés.
a bajo coste computacional y alta precisión. Nuestros experimentos también aportan
varias ideas: (i) los datos sintéticos de entrenamiento en profundidad son un excelente
sustituto de los datos reales; (ii) ampliar
El problema de aprendizaje con datos sintéticos variados es importante.
para alta precisión; y (iii) nuestro enfoque basado en partes generaliza mejor que
incluso un vecino más cercano exacto y oracular.

Trabajo relacionado. La estimación de la pose humana ha generado un
vasta literatura (revisada en [22, 29]). La reciente disponibilidad
El uso de cámaras de profundidad ha impulsado mayores avances [16, 19, 28].
Grest et al. [16] utilizan el punto más cercano iterado para rastrear una tonelada de
esqueleto de un tamaño y posición inicial conocidos. Anguelov et al.
[3] segmentar títeres en datos de escaneo de rango 3D en cabeza, extremidades,
torso y fondo usando imágenes de giro y un MRF. En
[44], Zhu y Fujimura construyen detectores heurísticos para
partes superiores del cuerpo (cabeza, torso, brazos) usando una relajación de
programación lineal, pero requieren una inicialización de la postura T al tamaño
el modelo. Siddiqui & Medioni [37] artesanía cabeza, mano,
y detectores de antebrazo, y muestra el modelo MCMC basado en datos
La adaptación supera la PIC. Kalogerakis et al. [18] clasificar y
segmentar vértices en una malla 3D completamente cerrada en diferentes
partes, pero no se ocupan de oclusiones y son sensibles a
topología de malla. Más similar a nuestro enfoque, Plagemann
et al. [28] construyen una malla 3D para encontrar puntos de interés geodésicos
extremos que se clasifican en 3 partes: cabeza, mano y
pie. Su método proporciona tanto una ubicación como una orientación.
estimación de estas partes, pero no distingue izquierda de
derecho y el uso de puntos de interés limita la elección de piezas.

También se han logrado avances utilizando cámaras de intensidad convencionales,
aunque normalmente con un costo computacional mucho mayor. Bregler y Malik [7]
rastrean a los humanos usando giros
y mapas exponenciales de una pose inicial conocida. Ioffe y
Forsyth [17] agrupa bordes paralelos como segmentos de cuerpo candidatos y poda
combinaciones de segmentos utilizando un clasificador proyectado. Mori y Malik [24]
utilizan el descriptor de contexto de forma para unir ejemplos. Ramanan y Forsyth

[31] encuentre segmentos corporales candidatos como pares de líneas paralelas,
agrupar apariencias en fotogramas. Shakhnarovich et al.
[33] estima la postura de la parte superior del cuerpo, interpolando posturas k-NN
coincide con hash sensible a parámetros. Agarwal y Triggs
[1] aprenda una regresión a partir de características de siluetas de imágenes
kernelizadas para posar. Sigal et al. [39] utilizan detectores de plantilla de apariencia
propia para propuestas de cabeza, parte superior de los brazos y parte inferior de las
piernas. Felzenszwalb & Huttenlocher [11] aplican una sesión fotográfica
estructuras para estimar la pose de manera eficiente. Navaratnam et al.
[25] utilizan las estadísticas marginales de datos sin etiquetar para mejorar la
estimación de pose. Urtasun y Darrel [41] propusieron una
Mezcla local de procesos gaussianos para hacer una regresión a la postura humana.
El contexto automático se utilizó en [40] para obtener una parte del cuerpo tosca.
El etiquetado, pero esto no estaba definido para localizar las articulaciones, y clasificar
cada cuadro tomó alrededor de 40 segundos. Rogez et al. [32]
entrenar bosques de decisiones aleatorias en una jerarquía de clases
definido en un toro de patrones cíclicos de movimiento humano y ángulos de cámara.
Wang y Popovic [42] rastrean una mano vestida con un
guante de color. Nuestro sistema podría verse como automáticamente
inferir los colores de un traje de color virtual desde una profundidad
imagen. Bourdev y Malik [6] presentan 'poselets' que forman
grupos apretados tanto en pose 3D como en apariencia de imagen 2D,
detectable mediante SVM.

2. Datos

La investigación sobre la estimación de pose a menudo se ha centrado en técnicas
para superar la falta de datos de entrenamiento [25], debido a dos problemas. En
primer lugar, la generación de imágenes de intensidad realista utilizando técnicas de
gráficos por computadora [33, 27, 26] se ve obstaculizada por la
Gran variabilidad de color y textura inducida por la ropa, el cabello,
y piel, lo que a menudo significa que los datos se reducen a siluetas 2D [1]. Aunque
las cámaras de profundidad reducen significativamente
esta dificultad, una variación considerable en el cuerpo y la vestimenta
la forma permanece. La segunda limitación es que el cuerpo sintético
Las imágenes de pose se alimentan necesariamente mediante captura de movimiento (mocap).
datos. Aunque existen técnicas para simular el movimiento humano.
(por ejemplo, [38]) todavía no producen el rango de volitivos
movimientos de un sujeto humano.

En esta sección revisamos las imágenes de profundidad y mostramos cómo
Usamos datos de mocap reales, reorientados a una variedad de modelos de
personajes básicos, para sintetizar un conjunto de datos grande y variado. Creemos
que este conjunto de datos avanzará considerablemente en el estado del arte.
tanto en escala como en variedad, y demostrar la importancia
de un conjunto de datos tan grande en nuestra evaluación.

2.1. Imágenes de profundidad

La tecnología de imágenes de profundidad ha avanzado dramáticamente
en los últimos años, alcanzando finalmente un precio al consumidor
punto con el lanzamiento de Kinect [21]. Píxeles en una imagen de profundidad
indicar la profundidad calibrada en la escena, en lugar de una medida
de intensidad o color. Empleamos la cámara Kinect que
Proporciona una imagen de 640x480 a 30 fotogramas por segundo con profundidad.
resolución de unos pocos centímetros.

Las cámaras de profundidad ofrecen varias ventajas sobre las tradicionales.
sensores de intensidad, trabajando en niveles bajos de luz, dando una estimación de
escala calibrada, siendo invariantes en color y textura, y
resolviendo ambigüedades de silueta en pose. Ellos también en gran medida

Figura 2. Datos sintéticos y reales. Pares de partes del cuerpo de imagen de profundidad y verdad del terreno. Observe la gran variedad en poses, formas, vestimenta y cultivos.

simplificar la tarea de resta de fondo que asumimos en este trabajo. Pero lo más importante para nuestro enfoque es que es sencillo sintetizar imágenes profundas realistas de personas y, por lo tanto, construir un gran conjunto de datos de entrenamiento de manera económica.

2.2. Datos de captura de movimiento

El cuerpo humano es capaz de adoptar una enorme variedad de posturas difíciles de simular. En cambio, capturamos una gran base de datos de captura de movimiento (mocap) de acciones humanas.
Nuestro objetivo era abarcar la amplia variedad de poses que la gente haría en un escenario de entretenimiento. La base de datos consta de aproximadamente 500.000 fotogramas en unos cientos de secuencias de conducir, bailar, patear, correr, navegar por menús, etc.

Esperamos que nuestro clasificador semilocal de partes del cuerpo se generalice un poco a poses invisibles. En particular, no necesitamos registrar todas las combinaciones posibles de los diferentes miembros; en la práctica, una amplia gama de posturas resulta suficiente. Además, no necesitamos registrar mocap con variación en la rotación alrededor del eje vertical, reflejo de izquierda a derecha, posición de la escena, forma y tamaño del cuerpo o pose de la cámara, todo lo cual se puede agregar (semi)automáticamente.

Dado que el clasificador no utiliza información temporal, sólo nos interesan las posturas estáticas y no el movimiento. A menudo, los cambios de pose de un fotograma de mocap al siguiente son tan pequeños que resultan insignificantes. Por lo tanto, descartamos muchas poses similares y redundantes de los datos iniciales de mocap usando la agrupación del 'vecino más lejano' [15] donde la distancia entre las poses j p1 y p2 se define como maxj p el máximo Eu- 1 −p 2 2, distancia clidiana sobre el cuerpo. articulaciones j. Usamos un umbral de 5 cm para quedarnos con aproximadamente 100k poses de manera que no haya dos poses que estén a menos de 5 cm.

Hemos descubierto que es necesario iterar el proceso de captura de movimiento, tomar muestras de nuestro modelo, entrenar el clasificador y probar la precisión de la predicción conjunta para refinar la base de datos mocap con regiones del espacio de pose que se habían omitido previamente. Nuestros primeros experimentos emplearon la base de datos CMU mocap [9] que dio resultados aceptables aunque cubría mucho menos espacio de pose.

2.3. Generación de datos sintéticos Construimos

un canal de renderizado aleatorio desde el cual podemos tomar muestras de imágenes de entrenamiento completamente etiquetadas. Nuestros objetivos al construir este canal eran dos: realismo y variedad. Para que el modelo aprendido funcione bien, las muestras deben parecerse mucho a imágenes de cámaras reales y contener una buena cobertura de las variaciones de apariencia que esperamos reconocer en el momento de la prueba. Si bien las variaciones de profundidad/escala y traducción se manejan explícitamente en nuestras funciones (ver más abajo), otras invariancias no se pueden codificar de manera eficiente. En cambio, aprendemos la invariancia de los datos con respecto a la postura de la cámara, la postura del cuerpo y el tamaño y la forma del cuerpo.

El proceso de síntesis primero toma muestras aleatorias de un conjunto de parámetros y luego utiliza técnicas estándar de gráficos por computadora para representar la profundidad y (ver más abajo) imágenes de partes del cuerpo a partir de mallas 3D con mapas de textura. El mocap se está reorientando a cada una de las 15 mallas base que abarcan una variedad de formas y tamaños corporales, utilizando [4]. Una ligera variación aleatoria adicional en altura y peso proporciona una cobertura adicional de las formas del cuerpo. Otros parámetros aleatorios incluyen el marco del mocap, la pose de la cámara, el ruido de la cámara, la ropa y el peinado.
Proporcionamos más detalles de estas variaciones en el material complementario. La Fig. 2 compara la salida variada de la tubería con imágenes de cámaras reales etiquetadas a mano.

3. Inferencia de partes del cuerpo y propuestas conjuntas

En esta sección describimos nuestra representación de partes intermedias del cuerpo, detallamos las características discriminativas de la imagen de profundidad, revisamos los bosques de decisión y su aplicación al reconocimiento de partes del cuerpo, y finalmente discutimos cómo se utiliza un algoritmo de búsqueda de modo para generar propuestas de posiciones conjuntas.

3.1. Etiquetado de partes del cuerpo

Una contribución clave de este trabajo es nuestra representación de la parte intermedia del cuerpo. Definimos varias etiquetas de partes del cuerpo localizadas que cubren densamente el cuerpo, según el código de colores en la Fig. 2.
Algunas de estas partes se definen para localizar directamente articulaciones esqueléticas particulares de interés, mientras que otras llenan los vacíos o podrían usarse en combinación para predecir otras articulaciones. Nuestra representación intermedia transforma el problema en uno que puede resolverse fácilmente mediante algoritmos de clasificación eficientes; mostramos en la sec. 4.3 que la penalización pagada por esta transformación es pequeña.

Las partes se especifican en un mapa de textura que se reorienta para revestir los distintos personajes durante el renderizado. Los pares de imágenes de profundidad y partes del cuerpo se utilizan como datos completamente etiquetados para aprender el clasificador (ver más abajo). Para los experimentos de este artículo, utilizamos 31 partes del cuerpo: LU/RU/LW/RW cabeza, cuello, hombro L/R , brazo LU/RU/LW/RW , codo L/R , muñeca L/ R , Mano derecha , torso LU/RU/LW/RW , pierna LU/RU/LW/RW , rodilla izquierda/derecha , tobillo izquierda/derecha , pie izquierdo/derecho (izquierdo, derecho, superior, inferior). Distinto

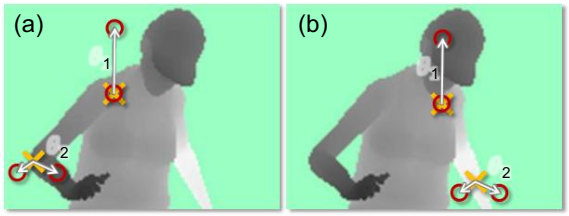

Figura 3. Características de la imagen de profundidad. Las cruces amarillas indican el píxel x que se está clasificando. Los círculos rojos indican los píxeles desplazados como se define en la ecuación. 1. En (a), las dos características de ejemplo dan una respuesta de gran diferencia de profundidad. En (b), las mismas dos características en ubicaciones de imágenes nuevas dan una respuesta mucho menor.

Las partes para izquierda y derecha permiten al clasificador eliminar la ambigüedad de los lados izquierdo y derecho del cuerpo.

Por supuesto, la definición precisa de estas partes podría cambiarse para adaptarse a una aplicación particular. Por ejemplo, en un escenario de seguimiento de la parte superior del cuerpo, todas las partes inferiores del cuerpo podrían fusionarse. Las piezas deben ser lo suficientemente pequeñas como para localizar con precisión las articulaciones del cuerpo, pero no demasiado numerosas como para desperdiciar la capacidad del clasificador.

3.2. Características de la imagen de profundidad

Empleamos funciones simples de comparación de profundidad, inspiradas por aquellos en [20]. En un píxel x dado, las características calculan

$$f_\theta(I, x) = d_I\left(x + \frac{u}{d_I(x)}\right) - d_I\left(x + \frac{v}{d_I(x)}\right), \quad (1)$$

donde $d_I(x)$ es la profundidad en el píxel x en la imagen I, y los parámetros $\theta = (u, v)$ describen desplazamientos u y v. La normalización 1 de los desplazamientos por garantiza que las características sean invariables en profundidad $d_I(x)$ ant: en un punto dado del cuerpo, se obtendrá un desplazamiento espacial fijo independientemente de que el píxel esté cerca o lejos de la cámara.
Por lo tanto, las características son invariantes de traducción 3D (efectos de perspectiva de módulo). Si un píxel desplazado se encuentra en el fondo o fuera de los límites de la imagen, la sonda de profundidad $d_I(x)$ recibe un valor constante positivo grande.

La figura 3 ilustra dos características en diferentes ubicaciones de píxeles x. La característica $f_{\theta 1}$ mira hacia arriba: Ec. 1 dará una respuesta positiva grande para los píxeles x cerca de la parte superior del cuerpo, pero un valor cercano a cero para los píxeles x que se encuentran más abajo en el cuerpo. En cambio, la característica $f_{\theta 2}$ puede ayudar a encontrar estructuras verticales delgadas como el brazo.

Individualmente, estas características proporcionan sólo una señal débil sobre a qué parte del cuerpo pertenece el píxel, pero en combinación en un bosque de decisiones son suficientes para eliminar con precisión todas las partes entrenadas. El diseño de estas funciones estuvo fuertemente motivado por su eficiencia computacional: no se necesita preprocesamiento; cada característica sólo necesita leer como máximo 3 píxeles de la imagen y realizar como máximo 5 operaciones aritméticas; y las funciones se pueden implementar directamente en la GPU. Dado un mayor presupuesto computacional, se podrían emplear características potencialmente más potentes basadas, por ejemplo, en integrales de profundidad sobre regiones, curvatura o descriptores locales, por ejemplo [5].

Bosques aleatorios

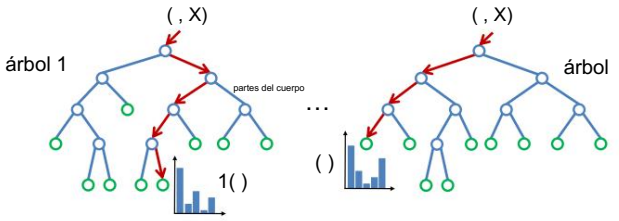

Figura 4. Bosques de decisión aleatoria. Un bosque es un conjunto de árboles. Cada árbol consta de nodos divididos (azul) y nodos de hoja (verde). Las flechas rojas indican los diferentes caminos que pueden tomar diferentes árboles para una entrada particular.

3.3. Bosques de decisión aleatoria

Los árboles y bosques de decisión aleatorios [35, 30, 2, 8] han demostrado ser clasificadores multiclase rápidos y eficaces para muchas tareas [20, 23, 36] y se pueden implementar de manera eficiente en la GPU [34]. Como se ilustra en la Fig. 4, un bosque es un conjunto de T árboles de decisión, cada uno de los cuales consta de nodos divididos y de hoja. Cada nodo dividido consta de una característica f_θ y un umbral τ.
Para clasificar el píxel x en la imagen I, se comienza en la raíz y se evalúa repetidamente la ecuación. 1, bifurcándose hacia la izquierda o hacia la derecha según la comparación con el umbral τ. En el nodo de hoja alcanzado en el árbol t, se almacena una distribución aprendida $P_t(c|I, x)$ sobre las etiquetas c de partes del cuerpo. Las distribuciones se promedian para todos los árboles del bosque para dar la clasificación final.

$$P(c|I, x) = \frac{1}{T}\sum_{t=1}^{T} P_t(c|I,x) . \quad (2)$$

Capacitación. Cada árbol se entrena con un conjunto diferente de imágenes sintetizadas aleatoriamente. Se elige un subconjunto aleatorio de 2000 píxeles de ejemplo de cada imagen para garantizar una distribución aproximadamente uniforme entre las partes del cuerpo. Cada árbol se entrena utilizando el siguiente algoritmo [20]:

1. Proponga aleatoriamente un conjunto de candidatos de división $\varphi = (\theta, \tau)$ (parámetros de característica θ y umbrales τ).
2. Divida el conjunto de ejemplos $Q = \{(I, x)\}$ a la izquierda y subconjuntos derechos por cada φ

$$Q_l(\varphi) = \{ (I, x) \mid f_\theta(y_o, x) < \tau \} \quad (3)$$

$$Q_r(\varphi) = Q \setminus Q_l(\varphi) \quad (4)$$

3. Calcule el φ que proporcione la mayor ganancia de información:

$$\varphi = \text{argmáx}_\varphi \; G(\varphi) \quad (5)$$

$$G(\varphi) = H(Q) - \sum_{s \in \{l,r\}} \frac{|Q_s(\varphi)|}{|P|} H(Q_s(\varphi)) \quad (6)$$

donde la entropía de Shannon H(Q) se calcula en el histograma normalizado de las etiquetas de partes del cuerpo $I_l(x)$ para todo $(I, x) \in Q$.

4. Si la mayor ganancia $G(\varphi)$ es suficiente y la profundidad en el árbol está por debajo de un máximo, entonces recurra para los subconjuntos izquierdo y derecho $Q_l(\varphi)$ y $Q_r(\varphi)$.

Inferencias de ejemplo

Figura 5. Ejemplos de inferencias. Sintético (fila superior); real (medio); modos de falla (abajo). Columna izquierda: verdad fundamental para una pose neutral como referencia. En cada ejemplo, vemos la imagen de profundidad, las etiquetas de las partes del cuerpo más probables inferidas y las propuestas conjuntas se muestran como vistas frontal, derecha y superior (superpuestas en una nube de puntos de profundidad). Solo se muestra la propuesta más confiable para cada articulación por encima de un umbral fijo y compartido.

• profundidad, mapa, frente/derecha/arriba
• distancias, recortes, ángulos de cámara, tamaño y forma del cuerpo (por ejemplo, niño
fallas con confianza •

Para mantener bajos los tiempos de entrenamiento empleamos un sistema distribuido. Los modos detectados se encuentran en la superficie del cuerpo. Cada pose, pequeño, delgado/gordo), implementación. Para probabilidad subyacente correcta, pose de profundidad 20 desde el modo 1 millón es devuelto a la escena mediante un imágenes sintéticas/reales/de fallas toma aproximadamente un día en un clúster de 1000 núcleos. z compensa ζc para producir una propuesta final de posición conjunta.

Este enfoque simple y eficiente funciona bien en la práctica. La banda- 3.4. Los anchos bc de las propuestas de posición conjunta , el umbral de probabilidad λ c y el reconocimiento de la parte del cuerpo de superficie a interior, como se describió anteriormente, infiere que el desplazamiento z por píxel ζc está optimizado por parte en una información de validación reservada. Esta información ahora debe agruparse en un conjunto de 5000 imágenes mediante una búsqueda en cuadrícula. (A modo de indicación, estos píxeles para generar propuestas confiables para las posiciones 3D dieron como resultado un ancho de banda medio de 0,065 m, umbral de probabilidad de las articulaciones esqueléticas. Estas propuestas son el resultado final de nuestro 0,14 y un desplazamiento z de 0,039 m). algoritmo, y podría ser utilizado por un algoritmo de seguimiento para autoinicializarse y recuperarse de una falla.

Una opción sencilla es acumular los centros de masa de probabilidad globales 3D para cada parte, utilizando la profundidad calibrada conocida. Sin embargo, los píxeles periféricos degradan gravemente la calidad de dicha estimación global. En su lugar, empleamos un enfoque de búsqueda de modo local basado en el desplazamiento medio [10] con un núcleo gaussiano ponderado.

Definimos un estimador de densidad por parte del cuerpo como

$$f_c(\hat{x}) \quad \sum_{yo=1}^{norte} \text{experiencia } w_{ic} - \left\| \frac{\hat{x} - \hat{x}_i \text{ antes de}}{\text{Cristo}} \right\|^2 , \qquad (7)$$

donde x̂ es una coordenada en el espacio mundial 3D, N es el número de píxeles de la imagen, wic es una ponderación de píxeles, x̂i es la reproyección del píxel de la imagen xi en el espacio mundial dada la profundidad dI (xi), y bc es un valor aprendido por -parte del ancho de banda. La ponderación de píxeles considera tanto la probabilidad inferida de la parte del cuerpo en el píxel como el área de superficie mundial del píxel:

$$w_{ic} = P(c|I, x_i) \cdot d_I(x_i)^2 . \qquad (8)$$

Esto garantiza que las estimaciones de densidad sean invariantes en profundidad y proporcionó una mejora pequeña pero significativa en la precisión de la predicción conjunta. Dependiendo de la definición de las partes del cuerpo, el P(c|I, x) posterior puede acumularse previamente en un pequeño conjunto de partes. Por ejemplo, en nuestros experimentos se fusionan las cuatro partes del cuerpo que cubren la cabeza para localizar la articulación de la cabeza.

El desplazamiento medio se utiliza para encontrar modos en esta densidad de manera eficiente. Todos los píxeles por encima de un umbral de probabilidad aprendido λ c se utilizan como puntos de partida para la parte c. Se proporciona una estimación de confianza final como la suma de los pesos de píxeles que alcanzan cada modo. Esto resultó más confiable que tomar la estimación de densidad modal.

4. Experimentos

En esta sección describimos los experimentos realizados para evaluar nuestro método. Mostramos resultados tanto cualitativos como cuantitativos en varios conjuntos de datos desafiantes, y los comparamos con enfoques del vecino más cercano y con el estado del arte [13]. Proporcionamos más resultados en el material complementario. A menos que se especifique lo contrario, los parámetros a continuación se establecieron como: 3 árboles, 20 de profundidad, 300 000 imágenes de entrenamiento por árbol, 2000 píxeles de ejemplo de entrenamiento por imagen, 2000 características candidatas θ y 50 umbrales candidatos τ por característica.

Datos de prueba. Utilizamos desafiantes imágenes sintéticas y de profundidad real para evaluar nuestro enfoque. Para nuestro conjunto de pruebas sintéticas, sintetizamos 5000 imágenes de profundidad, junto con las etiquetas reales de las partes del cuerpo y las posiciones de las articulaciones. Las poses de captura de movimiento originales utilizadas para generar estas imágenes se excluyen de los datos de entrenamiento. Nuestro conjunto de prueba real consta de 8808 fotogramas de imágenes con profundidad real de 15 sujetos diferentes, etiquetados a mano con partes densas del cuerpo y 7 posiciones de las articulaciones de la parte superior del cuerpo. También evaluamos los datos de profundidad real de [13]. Los resultados sugieren que los efectos observados en los datos sintéticos se reflejan en los datos reales y, además, que nuestro conjunto de pruebas sintéticas es, con diferencia, el "más difícil" debido a la extrema variabilidad en la postura y la forma del cuerpo. Para la mayoría de los experimentos, limitamos la rotación del usuario a ±120° tanto en los datos de entrenamiento como en los de prueba sintética, ya que el usuario está °) en nuestro entretenimiento principal frente a la cámara (escenario 0, aunque también evaluamos el escenario completo de 360°.

Métricas de error. Cuantificamos la precisión tanto de la clasificación como de la predicción conjunta. Para la clasificación, informamos la precisión promedio por clase, es decir, el promedio de la diagonal de la matriz de confusión entre la etiqueta de la parte de verdad fundamental y la etiqueta de la parte inferida más probable. Esta métrica pondera cada

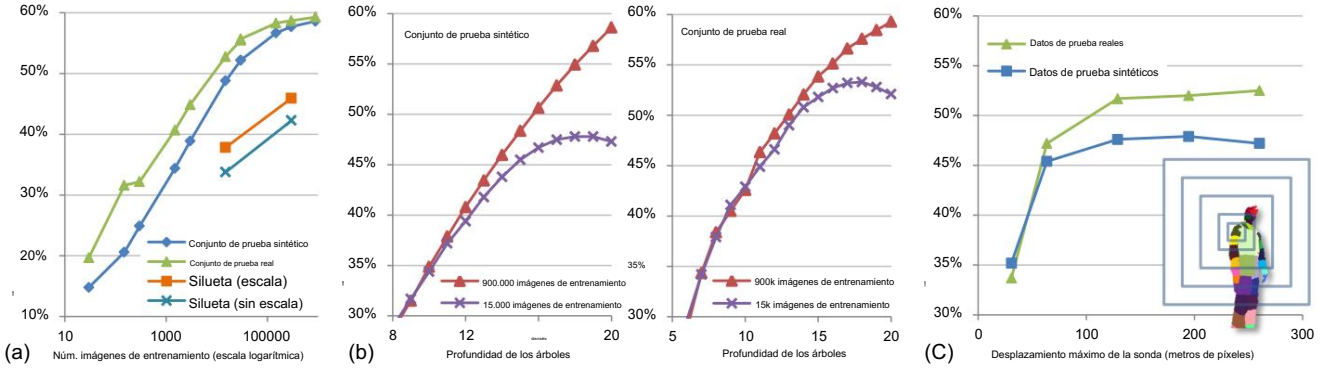

Figura 6. Parámetros de entrenamiento versus precisión de clasificación. (a) Número de imágenes de entrenamiento. (b) Profundidad de los árboles. (c) Desplazamiento máximo de la sonda.

partes del cuerpo por igual a pesar de sus diferentes tamaños, aunque las etiquetas incorrectas en los límites de las partes reducen los números absolutos.

Para propuestas conjuntas, generamos curvas de precisión de recuperación en función del umbral de confianza. Cuantificamos la precisión como precisión promedio por articulación, o precisión promedio promedio (mAP) sobre todas las articulaciones. La primera propuesta conjunta dentro de D metros de la posición real del terreno se toma como un verdadero positivo, mientras que otras propuestas también dentro de D metros cuentan como falsos positivos. Esto penaliza múltiples detecciones espurias cerca de la posición correcta, lo que podría ralentizar un algoritmo de seguimiento posterior. Cualquier propuesta conjunta fuera de los metros D también cuenta como falsos positivos. Tenga en cuenta que todas las propuestas (no sólo las más seguras) se cuentan en esta métrica. Las juntas invisibles en la imagen no se penalizan como falsos negativos. Establecemos D = 0,1 m a continuación, aproximadamente la precisión de la verdad sobre el terreno de los datos de prueba reales etiquetados a mano. La fuerte correlación entre la clasificación y la precisión de la predicción conjunta (véanse las curvas azules en las figuras 6(a) y 8(a)) sugiere que las tendencias observadas a continuación para una también se aplican para la otra.

4.1. Resultados cualitativos

La figura 5 muestra ejemplos de inferencias de nuestro algoritmo. Observe la alta precisión tanto de la clasificación como de la predicción conjunta a través de grandes variaciones en la pose del cuerpo y de la cámara, la profundidad de la escena, el recorte y el tamaño y la forma del cuerpo (por ejemplo, un niño pequeño versus un adulto pesado). La fila inferior muestra algunos modos de falla de la clasificación de partes del cuerpo. El primer ejemplo muestra una incapacidad para distinguir cambios sutiles en la imagen de profundidad, como los brazos cruzados. A menudo (como en el segundo y tercer ejemplo de falla) la parte del cuerpo más probable es incorrecta, pero todavía hay suficiente masa de probabilidad correcta en la distribución P(c|I, x) para que aún se pueda generar una propuesta precisa. El cuarto ejemplo muestra una incapacidad para generalizar bien a una pose invisible, pero la confianza bloquea malas propuestas, manteniendo una alta precisión a expensas de recordar.

Tenga en cuenta que no se utilizan restricciones temporales o cinemáticas (aparte de las implícitas en los datos de entrenamiento) para ninguno de nuestros resultados. A pesar de esto, los resultados por fotograma en las secuencias de vídeo del material complementario muestran que casi todas las articulaciones se predicen con precisión con una fluctuación notablemente pequeña.

4.2. Precisión de clasificación

Investigamos el efecto de varios parámetros de entrenamiento sobre la precisión de la clasificación. Las tendencias están altamente correlacionadas entre los conjuntos de prueba sintéticos y reales, y el conjunto de pruebas real parece consistentemente "más fácil" que el conjunto de pruebas sintético, probablemente debido a las poses menos variadas presentes.

Número de imágenes de entrenamiento. En la Fig. 6(a) mostramos cómo la precisión de la prueba aumenta aproximadamente de forma logarítmica con el número de imágenes de entrenamiento generadas aleatoriamente, aunque comienza a disminuir alrededor de 100k imágenes. Como se muestra a continuación, esta saturación probablemente se deba a la capacidad limitada del modelo de un bosque de decisión de 3 árboles y 20 profundidades.

Imágenes de silueta. También mostramos en la Fig. 6 (a) la calidad de nuestro enfoque en imágenes de silueta sintéticas, donde las características de la ecuación. 1 reciben una escala (como la profundidad media) o no (una profundidad fija y constante). Para la predicción conjunta correspondiente utilizando una métrica 2D con un umbral positivo verdadero de 10 píxeles, obtuvimos 0,539 mAP con escala y 0,465 mAP sin escala. Si bien es claramente una tarea más difícil debido a ambigüedades de profundidad, estos resultados sugieren la aplicabilidad de nuestro enfoque a otras modalidades de imágenes.

Profundidad de los árboles. La figura 6(b) muestra cómo la profundidad de los árboles afecta la precisión de la prueba utilizando imágenes de 15k o 900k. De todos los parámetros de entrenamiento, la profundidad parece tener el efecto más significativo ya que afecta directamente la capacidad del modelo del clasificador. Usando solo imágenes de 15k observamos un sobreajuste que comienza alrededor de la profundidad 17, pero el conjunto de entrenamiento ampliado de 900k evita esto. El gradiente de alta precisión en la profundidad 20 sugiere que se pueden lograr resultados aún mejores entrenando árboles aún más profundos, con un pequeño costo computacional de tiempo de ejecución adicional y una gran penalización de memoria adicional. De interés práctico es que, hasta aproximadamente la profundidad 10, el tamaño del conjunto de entrenamiento importa poco, lo que sugiere una estrategia de entrenamiento eficiente.

Desplazamiento máximo de la sonda. El rango de compensaciones de la sonda de profundidad permitidas durante el entrenamiento tiene un gran efecto en la precisión. Mostramos esto en la Fig. 6 (c) para imágenes de entrenamiento de 5k, donde 'desplazamiento máximo de la sonda' significa el máximo. Valor absoluto propuesto para las coordenadas x e y de u y v en la ecuación. 1. Los cuadros concéntricos de la derecha muestran los 5 valores máximos probados.

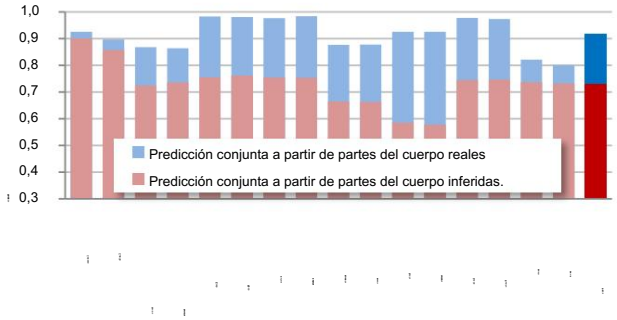

Figura 7. Precisión de predicción conjunta. Comparamos el rendimiento real de nuestro sistema (rojo) con el mejor resultado posible (azul) dadas las etiquetas reales de las partes del cuerpo.

conjuntos calibrados para un píxel del hombro izquierdo en esa imagen; el desplazamiento más grande cubre casi todo el cuerpo. (Recuerde que este desplazamiento máximo aumenta con la profundidad mundial del píxel). A medida que aumenta el desplazamiento máximo de la sonda, el clasificador puede utilizar más contexto espacial para tomar sus decisiones, aunque sin datos suficientes se corre el riesgo de sobreajustarse a este contexto. La precisión aumenta con el desplazamiento máximo de la sonda, aunque se nivela alrededor de 129 metros de píxeles.

4.3. Precisión de predicción conjunta En la

Fig. 7 mostramos resultados de precisión promedio en el conjunto de prueba sintético, logrando 0,731 mAP. Comparamos una configuración idealizada a la que se le asignan las etiquetas reales de las partes del cuerpo con la configuración real que utiliza partes del cuerpo inferidas. Si bien pagamos una pequeña penalización por utilizar nuestra representación de partes intermedias del cuerpo, para muchas articulaciones los resultados inferidos son muy precisos y cercanos a este límite superior. En el conjunto de prueba real, tenemos etiquetas reales para la cabeza, los hombros, los codos y las manos. Se logra un mAP de 0,984 en aquellas partes a las que se les dan las etiquetas de partes del cuerpo reales, mientras que se logra un mAP de 0,914 utilizando las partes del cuerpo inferidas. Como era de esperar, estos números son considerablemente más altos en este conjunto de pruebas más sencillo.

Comparación con el vecino más cercano. Para resaltar la necesidad de tratar el reconocimiento de pose en partes y calibrar la dificultad de nuestro conjunto de prueba para el lector, lo comparamos con dos variantes de coincidencia exacta de todo el cuerpo del vecino más cercano en la Fig. 8 (a). La primera variante, idealizada, hace coincidir el esqueleto de prueba de verdad del terreno con un conjunto de esqueletos ejemplares de entrenamiento con una alineación traslacional rígida óptima en el espacio mundial 3D. Por supuesto, en la práctica no se tiene acceso al esqueleto de prueba. Como ejemplo de un sistema realizable, la segunda variante utiliza la coincidencia de chaflanes [14] para comparar la imagen de prueba con los ejemplos de entrenamiento. Esto se calcula utilizando bordes de profundidad y 12 contenedores de orientación. Para facilitar la tarea de chaflán, descartamos cualquier imagen de prueba o entrenamiento recortada.

Alineamos imágenes utilizando el centro de masa 3D y descubrimos que una traducción rígida local adicional solo reducía la precisión.

Nuestro algoritmo, que reconoce en partes, generaliza mejor que incluso la coincidencia de esqueleto idealizada hasta alcanzar aproximadamente 150.000 imágenes de entrenamiento. Como se señaló anteriormente, nuestros resultados pueden ser aún mejores con árboles más profundos, pero ya hemos ro-

Infiere rápidamente las posiciones de las articulaciones del cuerpo en 3D y maneja de forma natural el recorte y la traducción. La velocidad de coincidencia de chaflanes del vecino más cercano también es drásticamente más lenta (2 fps) que nuestro algoritmo. Si bien la comparación jerárquica [14] es más rápida, todavía se necesitaría un conjunto de ejemplos masivo para lograr una precisión comparable.

Comparación con [13]. Los autores de [13] proporcionaron los datos de sus pruebas y los resultados para una comparación directa. Su algoritmo utiliza propuestas de partes del cuerpo de [28] y rastrea aún más el esqueleto con información cinemática y temporal. Sus datos provienen de una cámara de profundidad de tiempo de vuelo con características de ruido muy diferentes a las de nuestro sensor de luz estructurada. Sin ningún cambio en nuestros datos o algoritmo de entrenamiento, la Fig. 8 (b) muestra una precisión promedio de predicción conjunta considerablemente mejorada. Nuestro algoritmo también se ejecuta al menos 10 veces más rápido.

Rotaciones completas y múltiples personas. Para evaluar el escenario de rotación completa de 360°, entrenamos un bosque con imágenes de 900 000 que contienen rotaciones completas y lo probamos con imágenes sintéticas de rotación completa de 5 000 (con poses extendidas). A pesar del enorme aumento en la ambigüedad izquierda-derecha, nuestro sistema aún pudo lograr un mAP de 0,655, lo que indica que nuestro clasificador puede aprender con precisión las sutiles señales visuales que distinguen las posturas orientadas hacia adelante y hacia atrás. La incertidumbre residual de izquierda a derecha después de la clasificación puede propagarse naturalmente a un algoritmo de seguimiento a través de múltiples hipótesis. Nuestro enfoque puede proponer posiciones conjuntas para varias personas en la imagen, ya que el clasificador por píxel se generaliza bien incluso sin un entrenamiento explícito para este escenario. Los resultados se dan en la Fig. 1 y en el material complementario.

Propuestas más rápidas. También implementamos un enfoque alternativo más rápido para generar propuestas basado en una simple agrupación ascendente. Combinado con la clasificación de partes del cuerpo, esto se ejecuta a aproximadamente 200 fps en la GPU de Xbox, frente a aproximadamente 50 fps usando el cambio medio en una CPU de escritorio moderna de 8 núcleos. Dados los ahorros computacionales, los 0,677 mAP logrados en el conjunto de prueba sintético se comparan favorablemente con los 0,731 mAP del enfoque de cambio medio.

5. Discusión Hemos

visto cómo se pueden estimar propuestas precisas para las ubicaciones 3D de las articulaciones del cuerpo en tiempo súper real a partir de imágenes de profundidad única. Introdujimos el reconocimiento de partes del cuerpo como una representación intermedia para la estimación de la pose humana. El uso de un conjunto de entrenamiento sintético muy variado nos permitió entrenar bosques de decisión muy profundos utilizando características simples invariantes en profundidad sin sobreajuste, aprendiendo invariancia tanto para la pose como para la forma. La detección de modos en una función de densidad proporciona el conjunto final de propuestas de juntas 3D ponderadas por confianza. Nuestros resultados muestran una alta correlación entre los datos reales y sintéticos, y entre la clasificación intermedia y la precisión de la propuesta conjunta final. Hemos destacado la importancia de dividir todo el esqueleto en partes y mostrar una precisión de vanguardia en un conjunto de pruebas competitivo.

Como trabajo futuro, planeamos estudios adicionales de la variabilidad.

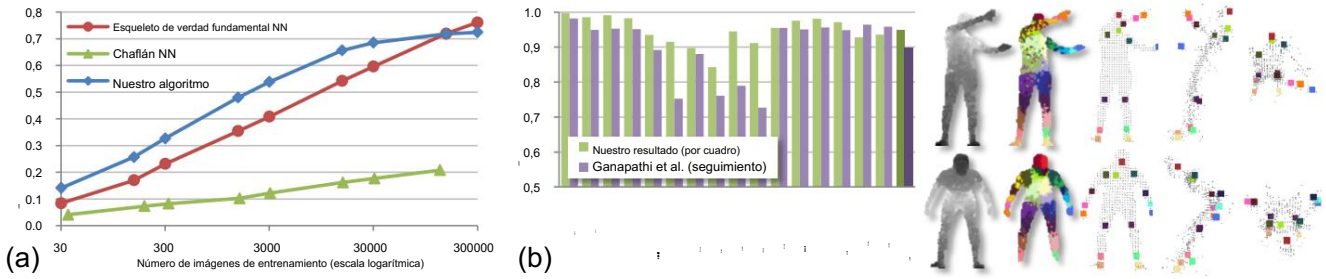

Figura 8. Comparaciones. (a) Comparación con la coincidencia del vecino más cercano. (b) Comparación con [13]. Incluso sin las limitaciones cinemáticas y temporales explotadas por [13], nuestro algoritmo es capaz de localizar con mayor precisión las articulaciones del cuerpo.

en los datos de origen de mocap, las propiedades del modelo generativo subyacente al proceso de síntesis y las definiciones de partes particulares. También es una cuestión abierta si existe un enfoque igualmente eficiente que pueda hacer retroceder directamente las posiciones conjuntas.

Quizás se podría utilizar una estimación global de variables latentes, como la orientación burda de la persona, para condicionar la inferencia de las partes del cuerpo y eliminar ambigüedades en las estimaciones de pose locales.

Agradecimientos. Agradecemos a los muchos ingenieros capacitados de Xbox, en particular a Robert Craig, Matt Bronder, Craig Peeper, Momin Al-Ghosien y Ryan Geiss, quienes construyeron el sistema de seguimiento Kinect a partir de esta investigación. También agradecemos a John Winn, Duncan Robertson, Antonio Criminisi, Shahram Izadi, Ollie Williams y Mihai Budiu por su ayuda y sus valiosas discusiones, y a Varun Ganapathi y Christian Plagemann por proporcionar los datos de sus pruebas.

Referencias

[1] A. Agarwal y B. Triggs. Pose humana 3D a partir de siluetas mediante regresión de vectores de relevancia. En Proc. CVPR, 2004. 2 [2] Y. Amit y D. Geman. Cuantización y reconocimiento de formas con árboles aleatorios. Neural Computation, 9(7):1545–1588, 1997. 4 [3] D. Anguelov, B. Taskar, V. Chatalbashev, D. Koller, D. Gupta y A. Ng. Aprendizaje discriminativo de campos aleatorios de Markov para la segmentación de datos de escaneo 3D. En Proc. CVPR, 2005. 2 [4] Autodesk MotionBuilder. 3 [5] S. Belongie, J. Malik y J. Puzicha. Coincidencia de formas y reconocimiento de objetos utilizando contextos de formas. Traducción IEEE. PAMI, 24, 2002. 4 [6] L. Bourdev y J. Malik. Poselets: detectores de partes del cuerpo entrenados utilizando anotaciones de poses humanas en 3D. En Proc. ICCV, 2009. 2 [7] C. Bregler y J. Malik. Seguimiento de personas con giros y mapas exponenciales. En Proc. CVPR, 1998. 1, 2

[8] L. Breiman. Bosques aleatorios. Mach. Learning, 45(1):5–32, 2001. 4 [9] Base de datos CMU Mocap. http://mocap.cs.cmu.edu/. 3 [10] D. Comaniciu y P. Meer. Cambio medio: un enfoque sólido hacia el análisis del espacio de características. Traducción IEEE. PAMI, 24(5), 2002. 1, 5

[11] P. Felzenszwalb y D. Huttenlocher. Estructuras pictóricas para el reconocimiento de objetos. IJCV, 61(1):55–79, enero de 2005. 2 [12] R. Fergus, P. Perona y A. Zisserman. Reconocimiento de clases de objetos mediante aprendizaje invariante de escala no supervisado. En Proc. CVPR, 2003. 1 [13] V. Ganapathi, C. Plagemann, D. Koller y S. Thrun. Captura de movimiento en tiempo real utilizando una única cámara de tiempo de vuelo. En Proc. CVPR, 2010. 1, 5, 7, 8 [14] D. Gavrila. Detección de peatones desde un vehículo en movimiento. En Proc. ECCV, junio de 2000. 7

[15] T. González. Agrupación para minimizar la máxima separación entre grupos. tancia. Teor. comp. Ciencia, 38, 1985. 3

[16] D. Grest, J. Woetzel y R. Koch. Estimación no lineal de la pose del cuerpo a partir de imágenes de profundidad. En En proceso. DAGM, 2005. 1, 2 [17] S. Ioffe y D. Forsyth. Métodos probabilísticos para encontrar personas. IJCV, 43(1):45–68, 2001. 2 [18] E. Kalogerakis, A. Hertzmann y K. Singh. Aprendizaje de segmentación y etiquetado de mallas 3D. Transmisión ACM. Gráficos, 29(3), 2010. 2

[19] S. Knoop, S. Vacek y R. Dillmann. Fusión de sensores para seguimiento del cuerpo humano en 3D con un modelo de cuerpo articulado en 3D. En Proc. ICRA, 2006. 1, 2

[20] V. Lepetit, P. Lagger y P. Fua. Árboles aleatorios para el reconocimiento de puntos clave en tiempo real. En Proc. CVPR, páginas 2:775–781, 2005.4

[21] Microsoft Corp. Redmond WA. Kinect para Xbox 360. 1, 2 [22] T. Moeslund, A. Hilton y V. Kruger. Un estudio de los avances en la captura y el análisis del movimiento humano basado en la visión. CVIU, 2006. 2 [23] F. Moosmann, B. Triggs y F. Jurie. Libros de códigos visuales discriminativos rápidos que utilizan bosques de agrupamiento aleatorio. En NIPS, 2006. 4 [24] G. Mori y J. Malik. Estimación de configuraciones del cuerpo humano mediante la coincidencia de contexto de forma. En Proc. ICCV, 2003. 2 [25] R. Navaratnam, AW Fitzgibbon y R. Cipolla. El modelo de variedad conjunta para regresión multivaluada semisupervisada. En Proc. ICCV, 2007. 2

[26] H. Ning, W. Xu, Y. Gong y TS Huang. Aprendizaje discriminativo de palabras visuales para la estimación de la pose humana en 3D. En Proc. CVPR, 2008. 2

[27] R. Okada y S. Soatto. Selección de características relevantes para la estimación y localización de la pose humana en imágenes desordenadas. En Proc. CEVC, 2008. 2

[28] C. Plagemann, V. Ganapathi, D. Koller y S. Thrun. Identificación y localización en tiempo real de partes del cuerpo a partir de imágenes de profundidad. En Proc. ICRA, 2010. 1, 2, 7 [29] R. Poppe. Análisis del movimiento humano basado en la visión: una descripción general. CVIU, 108, 2007. 2 [30] JR Quinlan. Inducción de árboles de decisión. Mach. Learn, 1986. 4 [31] D. Ramanan y D. Forsyth. Encontrar y rastrear personas del de abajo hacia arriba. En Proc. CVPR, 2003. 2

[32] G. Rogez, J. Rihan, S. Ramalingam, C. Orrite y P. Torr. Árboles aleatorios para la detección de poses humanas. En Proc. CVPR, 2008. 2 [33] G. Shakhnarovich, P. Viola y T. Darrell. Estimación de pose rápida con hash sensible a parámetros. En Proc. ICCV, 2003. 2 [34] T. Sharp. Implementación de árboles y bosques de decisión en una GPU. En Proc. ECCV, 2008. 1, 4 [35] B. Pastor. Una evaluación de un enfoque de árbol de decisión para la clasificación de imágenes. sificación. En IJCAI, 1983. 4

[36] J. Shotton, M. Johnson y R. Cipolla. Bosques de textos semánticos para categorización y segmentación de imágenes. En Proc. CVPR, 2008. 4 [37] M. Siddiqui y G. Medioni. Estimación de la postura humana desde un único punto de vista, sensor de alcance en tiempo real. En CVCG en CVPR, 2010. 1, 2 [38] H. Sidenbladh, M. Black y L. Sigal. Modelos probabilísticos implícitos del movimiento humano para síntesis y seguimiento. En ECCV, 2002. 2 [39] L. Sigal, S. Bhatia, S. Roth, M. Black y M. Isard. Seguimiento de personas con extremidades sueltas. En Proc. CVPR, 2004. 1, 2 [40] Z. Tu. Autocontexto y su aplicación a tareas de visión de alto nivel. En Proc. CVPR, 2008. 2

[41] R. Urtasun y T. Darrell. Regresión probabilística local para la inferencia de pose humana independiente de la actividad. En Proc. CVPR, 2008. 2 [42] R. Wang y J. Popovic. Seguimiento manual en tiempo real con un guante de color. En Proc. SIGGRAFO ACM, 2009. 1, 2

[43] J. Winn y J. Shotton. El diseño del campo aleatorio consistente para reconocer y segmentar objetos parcialmente ocluidos. En Proc. CVPR, 2006. 1

[44] Y. Zhu y K. Fujimura. Optimización restringida para la estimación de la pose humana a partir de secuencias de profundidad. En Proc. ACVC, 2007. 1, 2