SciVerse ScienceDirect
Disponible en www.sciencedirect.com

Revista Iberoamericana de Automática e Informática industrial 10 (2013) 170–177

www.elsevier.es/RIAI

Captura de movimiento y reconocimiento de actividades para múltiples personas mediante un enfoque bayesiano

A. Marcos*, D. Pizarro, M. Marrón, M. Mazo

Departamento de Electrónica, Carretera Madrid-Barcelona, Km 33,600. C.P.28871. Alcalá de Henares, Madrid, España

Resumen

Se presenta un método general para la detección, de forma no invasiva, de la postura corporal de varias personas a partir de la información capturada por múltiples cámaras. Se sigue una filosofía basada en el entrenamiento previo de un modelo articulado y posterior seguimiento. La principal aportación es la posibilidad de detectar varias personas simultáneamente. Se utiliza un modelo articulado para definir las posturas que puede adoptar una persona. Mediante bases de datos de captura de movimiento se selecciona un conjunto de clases o actividades predefinidas. El entrenamiento reduce la complejidad del modelo articulado a partir de técnicas no lineales de reducción de dimensionalidad. Así, las diferentes actividades de una persona quedan definidas de manera compacta por un conjunto de valores de baja dimensionalidad. Posteriormente, un filtro de partículas mixto (estados discretos y continuos) es utilizado para detectar la postura y el tipo de movimiento simultáneamente. Las hipótesis resultantes, seleccionadas automáticamente a partir de la distribución de partículas, son refinadas usando un optimizador no lineal que hace uso de funciones 'a priori' del tipo de movimiento entrenado. La propuesta se ha evaluado con un método simple pero estándar, basado en la comparación de volúmenes cilíndricos articulados con volúmenes del cuerpo humano, extraidos automáticamente a partir de las imágenes. Se consigue una precisión cercana a trabajos del estado del arte que no tienen en cuenta a más de una persona y ofrece un marco de trabajo flexible para futuras investigaciones.

Palabras Clave:
Visión artificial, aplicaciones de seguimiento, sistemas multidimensionales, visión estéreo.

1. Introducción

En los últimos años, una gran cantidad de trabajos han contribuido en la captura de movimiento no invasiva (sin marcas artificiales) de personas. En la mayoría de los casos la investigación se ha centrado en la detección de movimiento de una sola persona a partir de una red de cámaras. Hay pocos trabajos centrados en la detección de movimiento de múltiples personas. La capacidad de obtener simultáneamente la postura corporal de varias personas abre en gran medida el campo de aplicación de este tipo de algoritmos, como el reciente Kinect de Microsoft ha demostrado (Shotton et al. (2011)). Kinect utiliza una base de datos muy amplia, lo que permite asociar rápidamente formas tridimensionales y posición de articulaciones. Como contrapartida, es específico al sistema de observación con el que ha sido entrenado, en este caso, una cámara 2.5D. En otro trabajo reciente (Liu et al. (2011)) se propone una sistema capaz de detectar múltiples personas a partir de imágenes tomadas mediante una red de cámaras. En dicho trabajo resulta crucial la segmentación de las siluetas de las personas realizada en las cámaras, y aunque con resultados prometedores, está lejos de poder aplicarse en un entorno real con menos grado de control sobre la iluminación y el fondo.

Las técnicas de captura de movimiento pueden dividirse en tres categorías principales: métodos basados en aprendizaje, los cuales limitan el número de movimientos utilizables (Urtasun (2006)); métodos sin modelo, que no utilizan información "a priori" para modelar el cuerpo humano (Chu et al. (2003)); y métodos basados en modelo, que hacen corresponder un modelo humano creado "a priori" con la información observada, en cualquier tipo de movimiento (Bottino and Laurentini (2001), Gall et al. (2010a), Gall et al. (2009)). La propuesta realizada en este trabajo está a medio camino entre métodos basados en entrenamiento y métodos basados en modelo.

Existen básicamente tres variantes de sensado estudiadas en

*Autor en correspondencia.
Correos electrónicos: alvaro.marcos@depeca.uah.es (A. Marcos), pizarro@depeca.uah.es (D. Pizarro), marta@depeca.uah.es (M. Marrón), mazo@depeca.uah.es (M. Mazo)

http://dx.doi.org/10.1016/j.riai.2013.03.007

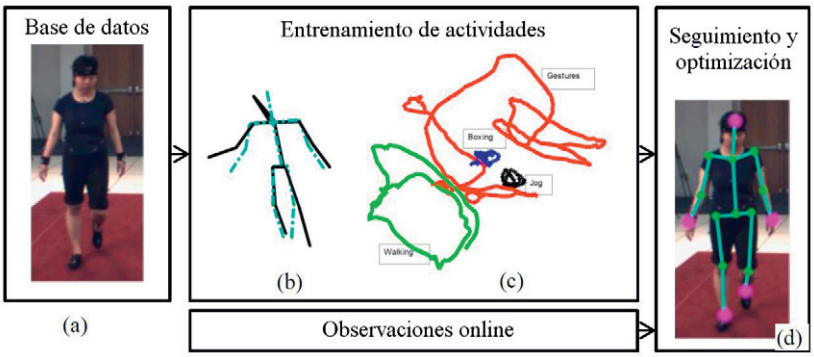

Figura 1: (a). Recolección de información a partir de la base de datos. (b) Alineación de esqueletos con SGPA (Stratified Generalized Procrustes Analysis). (c) Actividades representadas en el espacio latente. (d) Estimación final de las posturas

la literatura: el complicado caso monocular (Gall et al. (2010b)); los exitosos sensores 2.5D, como Kinect y las cámaras de tiempo de vuelo; y los métodos multicámara, donde varias cámaras permiten la utilización de propuestas basadas en Shape-from-Silhouette (SfS) para obtener volúmenes tridimensionales a partir de imágenes en múltiples cámaras. La propuesta presentada en este trabajo es un método general basado en la utilización de información 'a priori' de movimiento humano para seguir múltiples personas a partir de diversos sistemas de sensado. Para compararse con el estado del arte, gracias a la base de datos HumanEva (Sigal et al. (2010)), se ha elegido un método de reconstrucción tridimensional basado en la técnica de SfS.

El método propuesto se basa en una filosofía de entrenar para posteriormente hacer el seguimiento. A partir de bases de datos de captura de movimiento, se entrenan y etiquetan movimientos humanos realizados por varios individuos. Dado el alto número de dimensiones que se requieren para modelar el movimiento humano articulado, se utiliza una técnica de reducción de dimensionalidad para los movimientos entrenados. Hay estudios (Urtasun (2006)) que muestran que los movimientos humanos son intrínsecamente no lineales. Por lo tanto, se prefiere un mapeado no lineal a alternativas como Principal Component Analysis (PCA). De entre todas las posibilidades existentes, en este trabajo se propone el uso de GPLVM (Gaussian Process Latent Variable Model, Lawrence (2005)), debido a su popularidad y a que comparar la precisión de distintas técnicas de reducción de dimensionalidad se sale del objetivo de este trabajo. Desde la aparición de GPLVM, ha sido modificado, dando origen a algoritmos como SGPLVM (Grochow et al. (2004), GPDM (Wang et al. (2006)) o B-GPDM (Urtasun (2006)). Estas modificaciones no son relevantes de cara al propósito del trabajo propuesto.

La información obtenida con el entrenamiento se usa como conocimiento 'a priori' para un filtro Bayesiano utilizado para el seguimiento de múltiples personas. En particular se trata de un filtro de partículas de estado mixto (Isard and Blake (1998)). Utiliza estados discretos que identifican distintos movimientos, y estados continuos para parametrizar la postura (GPLVM codifica la posición y orientacion globales del cuerpo).

Finalmente, un algoritmo de optimización no lineal, que utiliza información estadística (mínimos cuadrados no lineales, Marquardt (1963)), entrega la postura final con precisión. El uso de un proceso de refinado no lineal permite una mejor detección de posturas que se alejan significativamente de aquellas usadas en el entrenamiento, utilizando la salida del filtro de partículas como punto de partida para la optimización. El esquema se muestra en la Figura 1. Las contribuciones principales del método propuesto son:

1. Estimación de la postura de múltiples personas mediante un único seguidor, gracias a un proceso de clasificación.
2. Proceso de seguimiento en dos etapas (primera aproximación mediante entrenamiento, proceso de generalización mediante optimización no lineal), que a su vez permite realizar reconocimiento de actividades

La robustez de esta propuesta se evalúa con una serie de experimentos en la sesión de resultados.

2. Resumen

La idea principal del método propuesto es utilizar el conocimiento del movimiento humano aprendido "a priori" para detectar distintas actividades y seguir múltiples personas. Esta información se consigue entrenando distintas actividades con un algoritmo de reducción de dimensionalidad (GPLVM). El resultado del entrenamiento es una serie de trayectorias relativamente simples en el espacio generado de dimensión reducida o espacio latente, como se muestra en la Figura 1. Estos datos, junto con el sensado online, se utilizan como entrada para un filtro de partículas. El cual retorna una aproximación de las posturas observadas (volúmenes obtenidos a partir de una técnica de SfS), basándose en la detección de hipótesis altamente probables. Estas posturas son refinadas con una técnica de optimización, que proporciona la salida final del algoritmo propuesto.

2.1. *Modelado de la postura*

La postura humana se parametriza utilizando un modelo articulado rígido con 20 puntos tridimensionales. Cada uno está se asocia a una de las N_{joints} del esqueleto humano elegidas. Es

decir, se utiliza un vector 60-dimensional, que se define como $\upsilon = \{x_{i_{joint}}, y_{i_{joint}}, z_{i_{joint}}\}_{i=1}^{N_{joints}}$.

Como se muestra en la Figura 2, para generar volúmenes humanos a partir de un vector de postura υ, se construye un modelo cilíndrico superpuesto al modelo articulado, para reflejar lo mejor posible las proporciones del cuerpo. Esta es una de las razones por las que se ha elegido parametrizar el modelo articulado con puntos 3D y no ángulos, ya que al utilizar cilindros para modelar segmentos corporales, se aporta una simetrá tal que uno de los ángulos de rotación del mismo no modifica la configuración de su superficie. Además, así se evitan ciertas ambiguedades que aparecen cuando los ángulos son cercanos a 0 y 360, y se obtiene una mejor representación en el espacio de baja dimensionalidad (Urtasun (2006)).

En este trabajo también se propone el uso de un método estándar de SfS para obtener volúmenes 3D de las personas presentes en la habitación, a partir de fusionar las siluetas resultantes de la segmentación de las imágenes obtenidas en las distintas cámaras. Ver Figura 9.

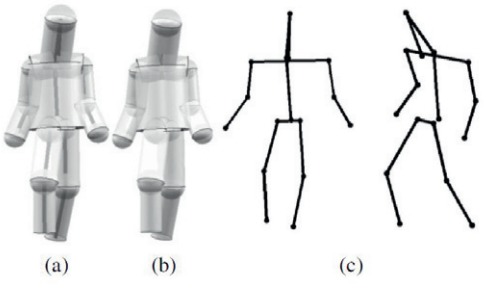

Figura 2: (a) Modelo cilíndrico utilizado con el esqueleto contenido. (b) Modelo cilíndrico (c) Diferentes vistas del modelo articulado de 60 dimensiones utilizado.

3. Proceso de entrenamiento

El objetivo del proceso de entrenamiento es encontrar una forma de generalizar, con un número pequeño de parámetros, la posición de las distintas articulaciones bajo una serie de movimientos etiquetados (andar, correr...). La posición precisa de las 20 articulaciones es conocida.

Una secuencia de entrenamiento está formada por un conjunto de posturas $\Upsilon = \{\upsilon_i\}_{i=1}^{N_T}$ de una persona, en una secuencia de N_T fotogramas. Se propone representar los vectores de postura en un espacio de pocas dimensiones conocido como espacio latente $\Lambda = \{\lambda_i\}_{i=1}^{N_T}$. Este mapeado, también conocido como función de regresión f, se define como:

$$\Upsilon = f(\Lambda, B) \tag{1}$$

donde B son los parámetros del mapeado, y representan los estadísticos de la dinámica del modelo. La función f puede ser lineal, como es el caso de PCA, o no lineal, como GPLVM. Como ya se ha mencionado, ya que los movimientos humanos son de naturaleza no lineal, se utiliza GPLVM para establecer el mapeado entre parejas $\{\lambda_i, \upsilon_i\}$.

Antes de expresar el conjunto de posturas Υ en el espacio latente, todo el conjunto se alinea mediante el método Procrustes Generalizado (Bartoli et al. (2010)) (Figura 1). Mediante el método de procrustes se eliminan los componentes de rotación y traslación de cada fotograma de la secuencia. Además, la postura media de la secuencia y_μ se resta a los datos. Ésta será posteriormente añadida para regenerar la postura completa a partir de un punto en el espacio latente.

La función f es entrenada iterativamente con GPVLM, refinando una inicialización de Λ que se ha obtenido mediante PCA. GPLVM está basado en PCA probabilístico (Lawrence (2005)) que añade una función kernel para permitir mapeados no lineales, a través de Procesos Gaussianos (Bernardo et al. (1992), Williams (1998)).

4. Proceso de seguimiento

Se utiliza un Filtro de Partículas (PF) de estado mixto para seguir múltiples posturas. A diferencia de un PF convencional, permite procesos donde conviven estados continuos y discretos en el vector de estado. Este método ha sido utilizado para conmutar automáticamente entre modelos (Isard and Blake (1998)). El método básico está descrito en Doucet (1997). Se define un espacio de estados extendido como se muestra:

$$X_t = (x_t, \delta_t), x_t \in \mathbf{R}^{N_M}, \delta_t \in \{1, ..., N_M\} \tag{2}$$

donde x_t es el vector de estados continuos convencional y δ_t es una variable discreta, que etiqueta al modelo asociado con el estado completo X_t. La matriz de probabilidades de transición $T = \{t_{ij}\}$ describe cómo de probable es ir desde el estado actual i hasta el estado j. El conjunto de partículas asociado con cada modo de probabilidad m en el instante t puede definirse como se muestra en la Ecuación 3:

$$\mathbf{S}_{m,t} = \{S_{mi,t}\}_{i=1}^{N_P} = \{(x_{\Lambda mi,t}, x_{Mmi,t}, x_{\Phi mi,t}, x_{mi,t})\}_{i=1}^{N_P} \tag{3}$$

donde el espacio de estados extendido propuesto se descompone como sigue: $x_{\Lambda mi} = (x_{\lambda 1}, \cdots, x_{\lambda Q})$ es un punto Q-dimensional en el espacio latente, $x_{Mmi} = (x_x, x_y, x_z)$ son coordenadas tridimensionales en el espacio de observación en el que la postura será situada, $x_{\Phi mi}$ es el ángulo de orientación de la postura en el eje longitudinal anatómico, y δ_{mi} es el estado discreto asociado a la partícula, que codifica el modelo de movimiento aprendido correspondiente. Finalmente, N_p es el número de partículas por modo. Por lo tanto, el número total de partículas es $N_{TP} = N_m N_p$. El algoritmo propuesto puede verse en la Figura 3, y se explica a continuación.

4.1. Inicialización

Se obtiene el conjunto de partículas inicial $\mathbf{S}_{m,t} = \{S_{mi,t}\}_{i=1}^{N_P}$ para cada modo m. Los distintos componentes del vector de estado se obtienen de la siguiente forma:

1. Las variables de estado discretas $\{\delta_{mi,t}\}_{i=1}^{N_P}$ se obtienen al muestrear los distintos movimientos que han sido aprendidos $(1, \cdots, N_M)$ según una distribución uniforme.

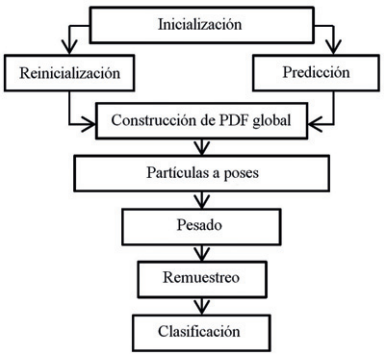

Figura 3: Resumen esquemático del proceso de seguimiento

2. Los puntos en el espacio latente $\{x_{\Lambda mi,t}\}_{i=1}^{N_P}$, se obtienen muestreando aleatoriamente del movimiento aprendido $\Lambda_{\delta_{mi,t}}$ asociado con la variable discreta, y luego añadiendo una dispersión $\Delta_{\Lambda i}$:

$$\begin{aligned} x_{\Lambda mi,t} &= (x_{\lambda 1}, \cdots, x_{\lambda Q}) + \Delta_{\Lambda i} \\ \Delta_{\Lambda i} &= (N(0, \kappa_\Lambda \sigma_{\lambda 1}), \cdots, N(0, \kappa_\Lambda \sigma_{\lambda Q})) \end{aligned} \quad (4)$$

donde $\{\sigma_{\lambda j}\}_{j=1}^{Q}$ es la desviación estándar del movimiento aprendido elegido, en cada dimensión j del espacio latente, y κ_Λ es una constate ajustada a mano.
3. Las coordenadas globales $\{x_{Mmi,t}\}_{i=1}^{N_P}$ se obtienen añadiendo una dispersión Δ_{Mi} al centroide de los distintos volúmenes SfS.
4. Finalmente, la orientación $\{x_{\Phi mi,t}\}_{i=1}^{N_P}$ se obtiene muestreando uniformemente el intervalo $\{0, 2\pi\}$ radianes.

4.2. Predicción y reinicialización

Se obtiene la predicción del conjunto de partículas:

$$\mathbf{S}_{m,t|t-1} = \{S_{mi,t|t-1}\}_{i=1}^{N_P'},$$

para cada modo m. El número de partículas para cada modo se define como:

$$N_P = N_P' + \alpha_K N_P + \alpha_R N_P \quad (5)$$

Donde α_K es una proporción del total de partículas disponibles para el modo, usada para reducir el efecto 'kidnapping', y α_R es otra porción utilizada para la reinicialización. Los distintos componentes del vector de estados se obtienen como sigue:

1. Los estados discretos $\{\delta_{mi,t}\}_{i=1}^{N_P}$ se muestrean de acuerdo con la matriz de transición T, y posteriormente la parte latente del vector de estado $\{x_{\Lambda mi,t}\}_{i=1}^{N_P}$ se calcula según la siguiente expresión:

$$x_{\Lambda mi,t} = \begin{cases} min(d(f^{-1}(x_{\Lambda mi,t-1}), \Lambda_{\delta_{mi,t-1}})) + \Delta_{\Lambda i} \\ \quad if \Lambda_{\delta_{mi,t}} = \Lambda_{\delta_{mi,t-1}} \\ x_{\lambda mi,t-1} + \Delta_{\lambda i} \quad otherwise \end{cases} \quad (6)$$

Es decir, si la actividad realizada no cambia, se añade dispersión al punto del entrenamiento más cercano a la postura actual. Si la actividad realizada cambia, solamente se añade dispersión.
2. Un porcentaje α_K del total N_P de partículas del modo se utiliza para distribuir $x_{\Lambda mi,t}$ uniformemente a lo largo de todo el espacio latente entrenado, como en el paso de inicialización. Las coordenadas globales $\{x_{Mmi,t}\}_{i=1}^{N_P}$ se obtienen dispersando el estado en el instante anteriror.
3. Se hace lo mismo para la orientación $\{x_{\Phi mi,t}\}_{i=1}^{N_P}$.

Finalmente, y conocido como paso de reinicialización, una proporción α_R del total de partículas disponible para el modo, es utilizado para buscar en todo el espacio de estados. Se utiliza el mismo método que en la inicialización, cada cierto número de iteraciones T_{Re}. Esto se hace para robustecer el algoritmo, permitiendo recuperarse mejor de pérdidas de seguimiento.

4.3. Formación de la Función Densidad de Probabilidad (PDF) y conversión de partículas

Los distintos modos de probabilidad se juntan antes de realizar los pasos de ponderación y remuestreo. La PDF se compone de la siguiente forma:

$$\mathbf{S}_{t|t-1}' = \{S_{m,t|t-1}\}_{m=1}^{N_m} \quad (7)$$

Una vez que $\mathbf{S}_{t|t-1}'$ está formada, cada partícula se convierte a una postura $y_{\rho i,t}$ reconstruida en el espacio de observaciones, aplicando el mapeado f^{-1} y añadiendo la media de la secuencia $y_{\mu\delta i}$ previamente restada. Por lo tanto, en este punto se tiene un conjunto de N_{TP} posturas en el espacio de observaciones, listas para aplicar el paso de ponderación.

4.4. Ponderación

Una función θ se aplica a cada postura $y_{\rho i,t}$ para obtener la misma postura parametrizada mediante ángulos, $y_{\Phi i,t}$, ya que será más útil para el proceso de optimización posterior:

$$y_{\Phi i,t} = \theta(y_{\rho i,t}) = (J_C, \Gamma) \quad (8)$$

donde $J_C = (j_{CX}, j_{CY}, j_{CZ})$ es la posición 3D de una de las articulaciones, y $\Gamma = \{\alpha_i.\beta_i\}_{i=1}^{N_j}$ es la parametrización en coordenadas esféricas de los N_j segmentos corporales. Un modelo cilíndrico Ψ es creado para recrear de forma aproximada las proporciones del volumen del cuerpo humano:

$$\Psi = \{\psi_i\}_{i=1}^{N_j} = \{\omega_i, \tau_i\}_{i=1}^{N_j} \quad (9)$$

Cada miembro τ_i define el radio del cilindro asociado a la parte del cuerpo, con longitud ω_i. Los valores de τ_i se establecen a mano, y las longitudes ω_i se obtienen mediante datos de captura de movimiento. El peso de cada partícula $w_{i,t}$ se calcula mediante la siguiente expresión:

$$w_{i,t} = \prod_{i=1}^{N_j} \xi_i \quad (10)$$

donde ξ_i es el porcentaje de llenado de cada segmento corporal. Es decir, el número de puntos observados $Y_{O,t}$ que caen dentro de las distancias definidas por Ψ, descritas como $\Omega(T_{O,t})$ son contadas y pesadas con $w_{O,t}$. Por lo tanto, cuanto más volumen observado haya dentro del modelo cilíndrico, y cuantas más cámaras hayan visto ese volumen, mayor peso tiene esa hipótesis. Analíticamente, el porcentaje de llenado por articulación i es:

$$\xi_i = \frac{\Omega(Y_{O,t})}{\frac{\pi\tau_i^2\omega_i}{\rho^3} + \frac{4\pi\tau_i^3}{3\rho^3}} \frac{\sum_{i=1}^{N_I} w_{Oi,t} | d(Y_{O,t}, \psi_i) < \tau_i}{N_I} \tag{11}$$

El primer término modela el porcentaje de llenado de cada cilindro. Como está definido como una distancia a segmento corporal, el cilindro tiene en realidad una semiesfera en cada extremo. Por ello, aparece el volumen de la esfera en el denominador. El segundo término mide cómo de probable fueron las observaciones que caen dentro del modelo. Está normalizado frente a la máxima posible N_I. Es decir, la probabilidad obtenida si todos los puntos que cayesen dentro del segmento corporal fuesen observados por todas las cámaras. Por lo tanto, ambos términos se encuentran dentro del intervalo [0, 1], lo que también se cumple para cada ξ_i y cada peso asociado $w_{i,t}$. Finalmente, se normaliza el conjunto de pesos.

4.5. *Remuestreo con asociación previa y normalización*

Con el sistema de pesos utilizado, es posible que un modo de probabilidad desapareza después de remuestrear si sus pesos son demasiado bajos respecto los de otros modos. Para evitar esto, se identifican y normalizan los distintos modos de probabilidad, como se muestra en la Figura 4

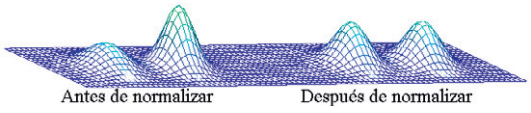

Figura 4: Modo probabilístico antes del remuestreo.

El paso de remuestreo es idéntico al Bootstrap: se reproducen las partículas con mayores pesos y se eliminan las de menor peso. Pero el número de partículas no permanece constante, ya que un cierto número volverá a ser insertado en el paso de reinicialización del próximo instante de tiempo. El conjunto de partículas obtenido es:

$$\mathbf{S}''_{t|t-1} = \{s''_{i,t|t-1}\}_{i=1}^{N_m(N_P'+\alpha_K N_P)} \tag{12}$$

4.6. *Asociación de partículas*

Se calcula el número de personas presentes en la escena utilizando un algoritmo de asociación lineal, con soporte para un número variable de grupos. Los datos de entrada para este algoritmo son la parte del vector de estados correspondiente a la posición global, $\{x_{Mi,t}\}_{i=1}^{N_m(N_P'+\alpha_K N_P)}$. Se asume por lo tanto que las distintas personas están los suficientemente alejadas unas de otras como para garantizar una clasificación existosa, lo que produce un conjunto de N_G grupos:

$$\{G_i\}_{i=1}^{N_G} = \{g_i, \{l_j\}_{j=1}^{N_{MGi}}\}_{i=1}^{N_G} \tag{13}$$

donde $g_i = (x_{gi}, y_{gi})$ es el centroide del grupo y l_j son los miembros asociados en el grupo. Las posturas finales en el espacio de observacion altamente dimensional $y_{\rho m,t}$ son obtenidas promediando las partículas remuestreadas que forman cada modo m.

En este punto hay N_G personas y N_m modos totales listos para el proceso de optimización.

5. Proceso de optimización

Además del proceso de seguimiento, se utiliza un proceso de optimización a partir de la salida del seguidor para mejorar las posturas finales. Se utiliza un algoritmo mínimos cuadrados no lineales para cada modo de las partículas, para minimizar el error ϵ:

$$\epsilon^2 = \sum_{i=1}^{N_m} \|1 - \{\xi_i\}_{i=i}^{N_j}\|^2 + \vartheta \|P - y_{pm,t}\|^2 \tag{14}$$

El primer término se utiliza para medir el porcentaje de llenado del modelo cilíndrico, ya que la variable ξ_i mide el porcentaje de llenado: es 0 si no hay ninguna parte de la reconstrucción dentro del modelo cilíndrico, y 1 si el modelo está completamente ocupado por la reconstrucción. Es decir, cuanto mejor se ajuste el modelo a las observaciones, ξ_i será más cercano a 1, y por lo tanto el término tenderá a 0 cuando es optimizado. Este término por lo tanto modela el mismo criterio utilizado al evaluar los pesos. El segundo término se utiliza para regularizar la postura final $y_{\rho m,t}$, y asegurar que no es demasiado distinta a la postura de salida P del seguidor, manteniendo propiedades inerciales. Cómo de distinto es permitido que sea se controla con el hiperparámetro ϑ, elegido empíricamente. N_m es el número de modos y N_j es el número de segmentos corporales. La Figura 5 muestra los efectos del refinamiento en las posturas:

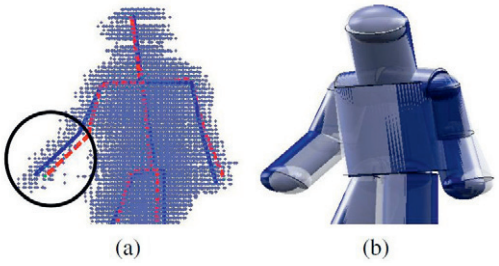

Figura 5: Efecto de la optimización en las posturas finales. (a) Punteado rojo: salida del seguidor. Azul continuo: postura optimizada. Puntos azules: vóxeles de SfS. (b) Gris: modelo volumétrico de la salida del seguidor. Azul oscuro: modelo volumétrico de la postura optimizada

6. Resultados

En esta sección se muestran los resultados de tres experimentos ideados para validar el concepto y aplicación del trabajo propuesto. Se pueden observar varios fotogramas de la salida del sistema en la Figura 11.

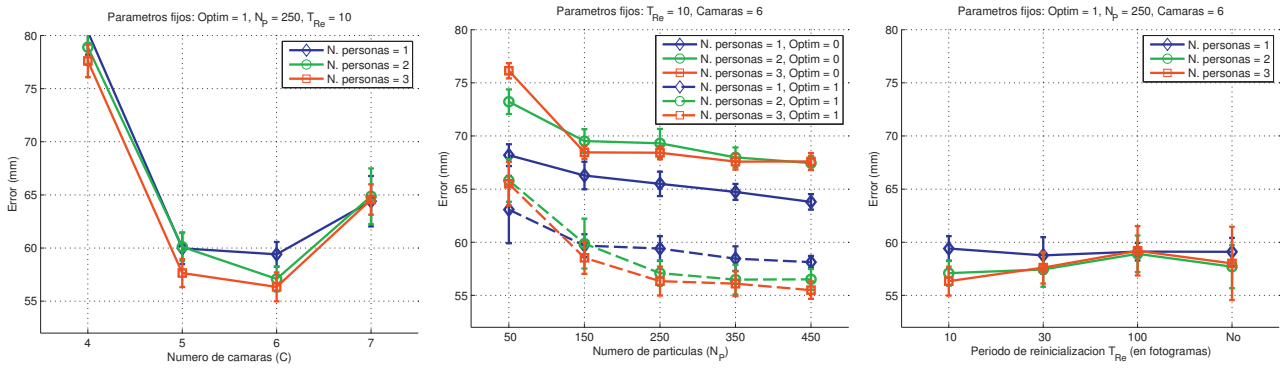

Figura 6: Resultados de precisión en relación con distintos parámetros. Izquierda: variando el número de cámaras. Centro: variando el número de partículas N_P, y habilitando o no el proceso de optimización. Derecha: variando el periodo de reinicialización T_{Re}. Resultados parametrizados para una, dos y tres personas en la escena.

1. Experimento A: El objetivo es conocer la cantidad mínima de articulaciones que es necesario observar para seguir la postura humana con un nivel de precisión aceptable.
2. Experimento B: Con el número de articulaciones mínima obtenida gracias al experimento A, se aplica el algoritmo de seguimiento propuesto para analizar su precisión en presencia de más de una persona.
3. Experimento C: La finalidad es evaluar el funcionamiento del método propuesto, en cuanto a precisión y reconocimiento de actividades, con un sistema de observación real, en este caso el SfS. Para ello se sigue un proceso de ajuste de los parámetros más relevantes del algoritmo propuesto.

Para la consecución de los objetivos ideados con estos experimentos, y además permitir una mejor comparación con el estado del arte, se utiliza la base de datos estándar de la comunidad, HumanEva. Se proporcionan imágenes de tres cámaras a color y cuatro en blanco y negro. Los parámetros del algoritmo que se han mantenido fijos a lo largo del proceso se muestran en la Tabla 1.

En el **experimento A** se utiliza una secuencia de 600 fotogramas, durante la que hay un cambio entre la acción "andar" y la acción "correr". Se mide el error de reconstrucción de la postura frente al número de articulaciones observadas. Se muestrea aleatoriamente en cada fotograma el número deseado de articulaciones, directamente desde el groundtruth (por lo tanto no se han utilizado SfSs en esta prueba), por lo que no hay error de posición intrínseco. Sin embargo, esto es válido para encontrar una relación entre error y cantidad de información disponible. Los resultados se muestran en la Figura 7.

Tabla 1: Parámetros utilizados para los experimentos

α_R	α_K	σ_{MXY}	σ_{MZ}	σ_Φ	T
250 %	15 %	150mm	20mm	0.45rad	uniforme

Se puede apreciar una reducción significativa el error al observar más de una articulación. A partir de dos articulaciones, el error decrece de forma mucho más progresiva. Esto se debe

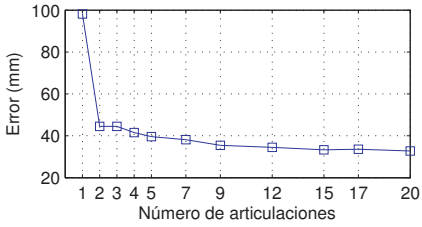

Figura 7: Error de seguimiento frente al número de articulaciones observadas.

a que la orientación $x_{\Phi mi}$ correcta se define con dos puntos tridimensionales, ya que se trata únicamente del ángulo en el eje longitudinal anatómico, como se explica en la sección 4. Una vez definido correctamente este ángulo, observar más articulaciones permite caracterizar de forma más detallada la postura. Dada la dificultad de observarlas de forma directa, consideramos que al utilizar 4 articulaciones (como podrían ser manos y pies, utilizando distancias geodésicas sobre SfS) se consigue un equilibrio aceptable entre cantidad de información a observar y precisión tridimensional media por articulación, siendo ésta cercana a 4 cm.

En el **experimento B** por tanto, se evalúa la capacidad del sistema propuesto para seguir a más de una persona simultáneamente. Para ello, se utiliza una secuencia de 600 fotogramas con dos personas en la escena (ver material suplementario). La primera persona comienza andando y pasa a correr en el fotograma 300. La segunda persona comienza corriendo y pasa a andar en el fotograma 340. El error medio, definido como distancia euclidea media entre las articulaciones del esqueleto reconstruido y el groundtruth, es **41 mm** al utilizar las 4 articulaciones que justifica el experimento A, a pesar de haber dos personas en la escena. El error temporal puede verse en la Figura 8.

Finalmente, para el **experimento C** se utiliza un sistema de observación basado en SfS, en el que cada vóxel no sólo está definido como "lleno" o "vacío", como es habitual, sino que también tiene un peso asociado al número de cámaras que ha observado ese punto. De esta forma, puede dar más confianza a partes de la escena que han sido vistas por más cámaras. Se ilustra en la Figura 9.

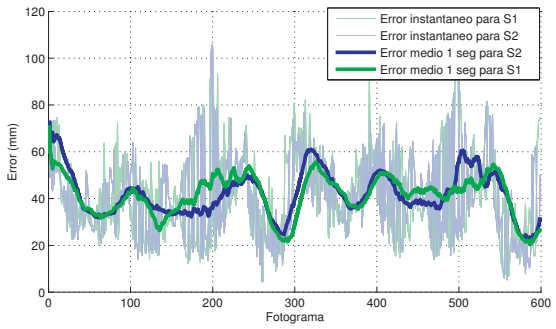

Figura 8: Error temporal de seguimiento en el experimento B. Se muestran resultados de error instantáneos y el error promedio utilizando una ventana temporal de 1 segundo, para dos sujetos (S1 y S2) distintos.

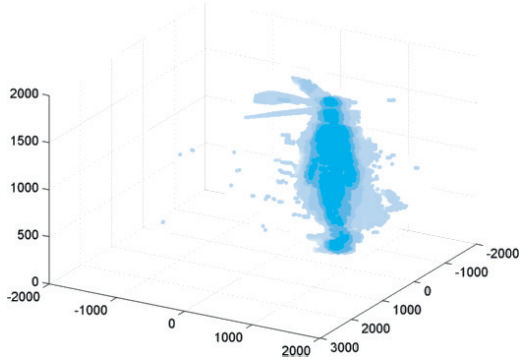

Figura 9: SfS con distinto número de cámaras. Colores más intensos denotan mayor número de cámaras. Se observa que con pocas cámaras errores de segmentación a lo largo de la escena aparecen sin filtrar, mientras que con muchas cámaras se omiten algunas zonas de las piernas, debido a la mala segmentación de los sensores en blanco y negro.

Se utilizan tres secuencias distintas de 300 fotogramas, cada una con un número distinto de personas (una, dos o tres), realizando las acciones "andar" y "correr", ya que debido a su similitud, sirven para poner a prueba la capacidad del sistema de reconocimiento de actividades. En la Figura 6 pueden verse los resultados de las pruebas realizadas para ajustar los parámetros del sistema. Cada experimento se ha repetido 10 veces para obtener una medida de confianza de la precisión obtenida en forma de desviación típica de la misma. En la Figura 6 izquierda se ha variado en número de cámaras utilizadas para la reconstrucción SfS. Puede verse que utilizar 6 cámaras arroja la mejor precisión. Esto es debido a que como se ha explicado anteriormente, en HumanEva hay tres cámaras a color y cuatro en blanco y negro, en las que la segmentación no es tan satisfactoria. Con 6 cámaras se consigue el equilibrio más favorable entre cantidad de información extraída de la escena e información descartada erróneamente debido a una mala segmentación.

Una vez fijado en número de cámaras a 6, en la Figura 6 centro puede verse el efecto de cambiar el número de partículas N_P, con y sin optimización. Se aprecia que el error decrece monótonamente al aumentar N_P, como es habitual en un filtro de partículas. Al introducir el proceso de optimización fina, se consigue reducir de forma sistemática el error, debido a una mejor generalización. Consideramos que $N_P = 250$ proporciona un equilibrio aceptable entre precisión y tiempo de ejecución.

Por último, en la Figura 6 derecha se muestra el efecto de variar el número de iteraciones tras el cual se introduce una nueva etapa de reinicialización (T_{Re}, apartado 4.2). Puede verse que no hay una tendencia clara en cuanto a precisión media. Sin embargo, periodos cortos consiguen reducir la varianza de la misma, ya que se guían las partículas con más frecuencia hacia la zona correcta del espacio de estados, reduciendo pérdidas de seguimiento y efectos de kidnapping.

En cuanto al sistema de reconocimiento de actividades, hemos apreciado una alta precisión con los parámetros fijados (6 cámaras, $N_P = 250$, optimización habilitada, $T_{Re} = 10$), por lo que se muestran los resultados en la Figura 10 respecto al número de cámaras y el uso o no de la optimización fina, parámetros con los cuales aparecen las mayores variaciones.

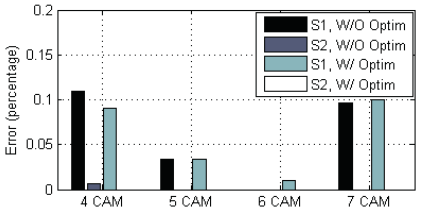

Figura 10: Resultado del reconocimiento de acciones para dos sujetos realizando acciones distintas, con y sin optimización, variando el número de cámaras.

7. Conclusión

Se ha presentado un nuevo método para obtener de forma simultánea la postura e información de actividad de múltiples personas. Se utiliza un seguidor Bayesiano para realizar el seguimiento de posturas de forma rápida y eficiente, con la ayuda de un espacio de baja dimensionalidad como GPLVM. Esta estimación es mejorada con un algoritmo de optimización, que dota al sistema de una capacidad extra de generalización. El error de precisión se corresponde con el equivalente en el estado del arte para modelos humanos de este tipo (a partir de primitivas básicas como cilindros, en lugar de realizar un escaneado láser a medida de cada sujeto). El reconocimiento de actividades se ha mostrado efectivo, distinguiendo con éxito movimientos similares.

En cuanto a limitaciones del método propuesto, cabe destacar que la interacción entre personas no está soportada robustamente, a no ser que se hagan cambios específicos en el sistema de observación, como un método de segmentación diferente. Además, el paso de reinicialización del seguidor debería ser revisado para tener mejor en cuenta las observaciones. Se van a abordar estos puntos en trabajos futuros.

English Summary

A bayesian approach to markerless motion capture and activity recognition of multiple people

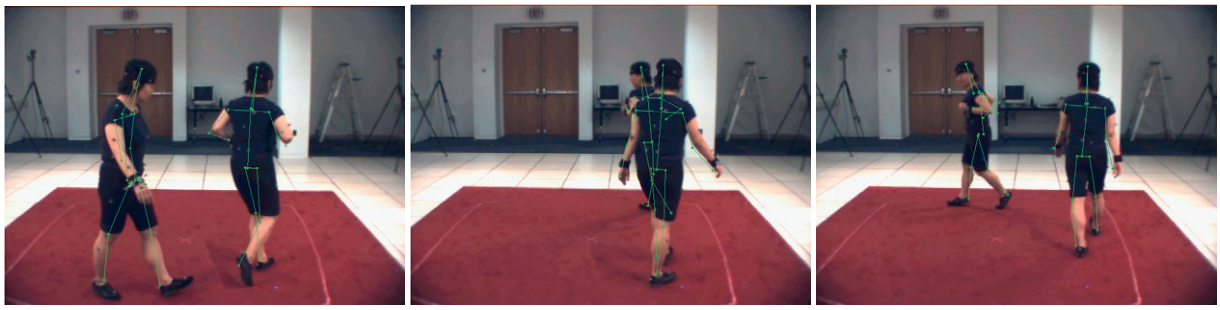

Figura 11: Fotograma del experimento 3. Las imágenes de HumanEva se han alterado para representar dos personas simultánemente. Los esqueletos reconstruidos se han superpuesto en las imágenes.

Abstract

This work presents a general framework for tracking simultaneously the body posturas of multiple people from non-intrusive visual sensors. The method is based on a training-then-tracking philosophy, with the main addition of being able to handle more than just one person. We train the body postura from labelled motion capture datasets. The training process is based on popular non-linear dimensionality reduction techniques. Then, a mixed, discrete and continuous state particle filter is used to simultaneously detect the postura and the kind of motion performed by each of the human bodies. The resulting hypotheses, automatically selected from the particle distribution, are then refined using non-linear optimization methods with statistical priors. The whole framework is tested using a simple but standard method based on comparing articulated cylindrical models with SfS volumes, taken from several cameras. Our accuracy in public available datasets is near to the state-of-the-art works that do not take into account multiple people in the problem.

Keywords:

Computer vision, tracking applications, multidimensional systems, stereo vision

Agradecimientos

Este trabajo ha sido parcialmente respaldado por el Ministerio de Ciencia e Innovación Español bajo los proyectos VISNU (ref. TIN2009-08984) y SDTEAM-UAH (ref. TIN2008-06856-C05-05).

Referencias

Bartoli, A., Pizarro, D., Loog, M., 2010. Stratified Generalized Procrustes Analysis. British Machine Vision Conference.

Bernardo, J., Berger, J., Dawid, A., Smith, A., 1992. Some Bayesian Numerical Analysis. Bayesian Statistics, Vol. 4.

Bottino, A., Laurentini, A., 2001. A silhouette based technique for the reconstruction of human movement. Computer Vision and Image Understanding 83 (1), 79–95.

Chu, C., Jenkins, O., Mataric, M., 2003. Markerless kinematic model and motion capture from volume sequences.

Doucet, A., 1997. Monte-Carlo methods for bayesian estimation of Hidden-Markov models. Application to Radiation Signal. Ph.D. thesis, University of Paris-Sud, France.

Gall, J., Rosenhahn, B., Brox, T., Seidel, H., 2010a. Optimization and filtering for human motion capture. International journal of computer vision 87 (1), 75–92.

Gall, J., Stoll, C., De Aguiar, E., Theobalt, C., Rosenhahn, B., Seidel, H., 2009. Motion capture using joint skeleton tracking and surface estimation.

Gall, J., Yao, A., Van Gool, L., 2010b. 2d action recognition serves 3d human pose estimation. Computer Vision–ECCV 2010, 425–438.

Grochow, K., Martin, S., Hertzmann, A., Popović, Z., 2004. Style-based inverse kinematics. In: ACM Transactions on Graphics (TOG). Vol. 23. ACM, pp. 522–531.

Isard, M., Blake, A., 1998. A mixed-state condensation tracker with automatic model-switching. In: Computer Vision, 1998. Sixth International Conference on. IEEE, pp. 107–112.

Lawrence, N., 2005. Probabilistic non-linear principal component analysis with Gaussian process latent variable models. The Journal of Machine Learning Research 6, 1783–1816.

Liu, Y., Stoll, C., Gall, J., Seidel, H., Theobalt, C., 2011. Markerless motion capture of interacting characters using multi-view image segmentation. Computer Vision and Pattern Recognition.

Marquardt, D., 1963. An algorithm for least-squares estimation of nonlinear parameters. Journal of the society for Industrial and Applied Mathematics 11 (2), 431–441.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A., 2011. Real-time human pose recognition in parts from single depth images. In: In CVPR.

Sigal, L., Balan, A., Black, M., 2010. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. International Journal of Computer Vision 87 (1), 4–27.

Urtasun, R., 2006. Motion Models for Robust 3D Human Body Tracking. Ph.D. thesis, ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE.

Wang, J., Fleet, D., Hertzmann, A., 2006. Gaussian process dynamical models. Advances in neural information processing systems 18, 1441.

Williams, C., 1998. Prediction with Gaussian processes: From linear regression to linear prediction and beyond. NATO ASI SERIES D BEHAVIOURAL AND SOCIAL SCIENCES 89, 599–621.