Actas de las XXXIX Jornadas de Automática, Badajoz, 5-7 de Septiembre de 2018
https://doi.org/10.17979/spudc.9788497497565.0967

RE-IDENTIFICACION DE PERSONAS MEDIANTE LA DISTANCIA DE MAHALANOBIS

María J. Gómez-Silva, José M. Armingol, Arturo de la Escalera
Laboratorio de Sistemas Inteligentes (LSI), Universidad Carlos III de Madrid, Avda. de la Universidad, 30, 28911 Leganés, Madrid, España
{magomezs, armingol, escalera }@ing.uc3m.es

Resumen

La re-identificación de una persona requiere del aprendizaje de una distancia métrica capaz de comparar dos imágenes y decidir si pertenecen o no a la misma persona. La automatización de esta tarea, para su aplicación en videovigilancia inteligente, plantea un gran reto debido a la presencia de personas con una apariencia similar. Por ello, es necesario el aprendizaje de características discriminativas, y de una métrica que las combine apropiadamente. Sin embargo, las variaciones de iluminación, perspectiva, fondo, resolución o escala entre dos imágenes de una misma persona, capturada desde vistas diferentes, hacen que su apariencia varíe, dificultando su re-identificación. Este artículo propone la codificación de las trasformaciones entre las vistas, en una matriz de Mahalanobis, cuya estimación ha sido integrada en el aprendizaje de las características discriminativas, de modo que estas últimas puedan reflejar las disimilitudes principalmente debidas a cambios de apariencia y no de punto de vista. Esta estimación ha sido implementada como una nueva capa de una red neuronal convolucional profunda, que ha sido entrenada y evaluada con la base de datos PRID2011.

Palabras clave: Re-Identificación de Personas, Matriz de Mahalanobis, Red Neuronal Convolucional.

1 INTRODUCCIÓN

La re-identificación de una persona consiste en su reconocimiento en imágenes capturadas en distintos instantes, desde dos cámaras con campos de visión no solapados. Existen dos tipos de re-identificación en función del número de imágenes de una cierta persona que son capturadas desde cada cámara, la basada en múltiples capturas, (multi-shot), [2], o en una única imagen desde cada vista (single-shot), [14]. Éste último es el tipo de reconocimiento abordado en este trabajo. La resolución de esta tarea ha despertado un gran interés investigador ya que permite el seguimiento y análisis del comportamiento de un individuo desde múltiples puntos de monitorización. Sin embargo, esto entraña un gran reto debido a la presencia de variaciones intra-clase y ambigüedades inter-clase. Las primeras se refieren a las variaciones en la apariencia de una persona causadas por cambios de iluminación, pose, perspectiva, resolución, o escala, y las segundas ocurren cuando distintas personas presentan un aspecto similar.

Con el objetivo de resolver estos problemas, una gran cantidad de trabajos se han dedicado al diseño de características que representen los atributos más discriminativos y significativos de la apariencia de un individuo [11]. Además, el desarrollo de técnicas de aprendizaje profundo (deep learning) ha impulsado nuevos métodos basados en el entrenamiento de redes neuronales convolucionales (CNN, convolutional neural networks) para extraer características de alto nivel a partir de los píxeles de una imagen. El aprendizaje de descriptores para re-identificación ha sido comúnmente realizado mediante redes siamesas, [1], que consisten en dos CNNs que comparten los mismos parámetros y son unidas por una distancia métrica y una función de coste, que lleva a la red a discriminar entre parejas de imágenes de una misma persona o de dos distintas. En este trabajo, para extraer las características más representativas, se ha empleado el marco de entrenamiento descrito en [5], donde se utiliza una arquitectura siamesa en la que cada rama aprende características para nueve partes del cuerpo, previamente extraídas de cada imagen.

Por otra parte, otros métodos han sido desarrollados para mejorar no la capacidad representativa de las características, sino la combinación y comparación de las mismas, a partir de una distancia métrica. En [8], es usada la distancia Euclídea. Sin embargo, recientemente, la investigación se ha centrado en la búsqueda de la métrica óptima. En [12], se propone un método basado en la ponderación de la importancia de las características según distintos prototipos, o grupos de población, (PSFI, Prototype-Sensitive Feature Importance). Por el contrario, el método aquí presentado sigue un enfoque conocido como Importancia Global de las Características (GFI, Global Feature Importance), con el que se aprende una ponderación global, constituida por un array de pesos que definen la estabilidad de cada característica a través de ambas vistas. Este enfoque de GFI, ha sido

ampliamente extendido y adoptado por algunos métodos como Ranking Support Vector Machines (Rank-SVM), [15], Probabilistic Relative Distance Comparison (PRDC), [19], o algoritmos de aprendizaje de la métrica (Metric Learning), tales como, Linear Discriminant Analysis (LDA), [3], y Logistic Discriminant Metric Learning (LDML), [6].

Rank-SVM, [15], aprende un peso independiente para cada característica, y PRDC, [19], una matriz ortogonal que codifica la importancia global de cada característica. En lugar de esto, el aprendizaje de la distancia de Mahalanobis optimiza una matriz que relaciona las características extraídas para una de las vistas con las de la otra, explotando la estructura de los datos bajo la asunción de que ambos grupos presentan la misma distribución probabilística. Esta distancia ha sido la empleada, en este trabajo, para codificar las transiciones entre vistas, reduciendo el efecto de las variaciones de apariencia debidas a ellas.

Algunos de los algoritmos mencionados evalúan la importancia discriminativa de cada una de las características extraídas, [12], y otros, como [3, 6, 16], aprenden una distancia que incorpore de forma implícita las transformaciones entre vistas. En este trabajo, ambas funciones son realizadas mediante la estimación de la matriz de Mahalanobis a partir del análisis de las características calculadas y a su integración en el proceso de aprendizaje de las mismas. La mayor parte de los algoritmos de aprendizaje de una métrica, optimizan una función lineal para ponderar adecuadamente la diferencia absoluta entre las características de dos imágenes, después de que éstas hayan sido calculadas para una base de datos. Este es el caso de LDA [3], cuyo objetivo es estimar una clasificación lineal de características para separar dos clases distintas. Este método es una generalización del método lineal discriminante de Fisher [18], y LDML, [6], sigue el mismo enfoque desde una perspectiva probabilística. Por el contrario, este artículo propone el aprendizaje simultáneo, tanto de los descriptores, como de la distancia óptima para compararlos.

La principales contribuciones de este trabajo son: i) el diseño de un método de análisis discriminativo de la estructura de los datos, para estimar la matriz de Mahalanobis; ii) estudio de las medidas precisas para la integración de tal análisis, tanto en la arquitectura neural como en su proceso de entrenamiento; iii) empleo del método de la media móvil exponencialmente ponderada (EWMA, Exponentially Weighted Moving Average) en cada iteración del aprendizaje para conservar la contribución de los datos analizados en iteraciones anteriores, además de su comparación con el método basado en colas de almacenamiento, presentado en [5], que requiere un gasto mayor de memoria computacional.

Se han realizado experimentos sobre la base de datos PRID2011, [7], una de las más desafiantes y usadas para la evaluación de algoritmos de re-identificación. Ésta presenta dos conjuntos de imágenes, cada uno capturado desde una cámara, que contienen un ejemplo de cada individuo a re-identificar. Esto ha permitido probar la capacidad de la matriz de Mahalanobis aprendida para codificar las variaciones entre dos vistas, ofreciendo excelentes resultados.

El resto del artículo está estructurado de la siguiente manera: la Sección 2 y 3 describen el método de re-identificación propuesto, y su entrenamiento, respectivamente. Los resultados experimentales son presentados en la Sección 4 y, por último, algunas conclusiones son expuestas en la Sección 5.

2 ALGORITMO PROPUESTO

La re-identificación basada en una única captura (single-shot) tiene como objetivo identificar a la persona representada en una imagen de prueba (probe image), capturada desde una de las cámaras, entre una galería de imágenes capturadas desde la otra cámara (gallery images). Para ello, cada imagen de prueba es comparada con todas las de la galería, mediante una distancia métrica, como describe la siguiente sección.

2.1 DISTANCIA MÉTRICA

La comparación de las imágenes no se realiza directamente a partir de los valores de sus píxeles. En su lugar, es calculado un descriptor para cada imagen, i.e. su representación en el espacio de las características, y es medida la distancia entre ellos. La trasformación de cada imagen, I, en su correspondiente representación, $F(I)$, es realizada por el modelo de red neuronal propuesto en [5] y mostrado en la figura 1. En esta red, una primera capa extrae nueve imágenes correspondientes a las partes del cuerpo de la persona representada en la imagen de entrada. Las partes extraídas son cabeza, brazos y piernas, dividiendo las cuatro extremidades en parte superior e inferior. Posteriormente, nueve redes convolucionales, una por cada parte, extraen características, que finalmente son agrupadas y ponderadas para generar un único descriptor de la persona, formado por un vector de características.

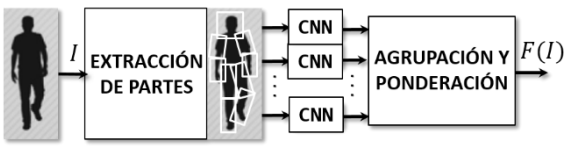

Figura 1: Modelo de red neuronal para calcular el vector de características, $F(I)$, de una imágen.

Finalmente, los descriptores son comparados por la distancia de Mahalanobis al cuadrado, $d_M^2\left(F(I^a), F(I^b)\right)$, de acuerdo con la ecuación 1, donde *M* es la matriz de Mahalanobis, aprendida con el método descrito en la siguiente sección.

$$d_M^2\left(F(I^a), F(I^b)\right) = \left(F(I^a) - F(I^b)\right)^T M\left(F(I^a) - F(I^b)\right) \quad (1)$$

2.2 ESTIMACIÓN DE LA MATRIZ DE MAHALANOBIS

La matriz de Mahalanobis es estimada mediante un método discriminativo de análisis de la estructura de los datos. Los datos analizados son las características aprendidas para distintos pares de imágenes, $(F(I^a), F(I^b))$. Demonimanos par positivo al formado por imágenes de una misma persona, y negativo al formado por representaciones de distintos individuos. Los pares son etiquetados con la función *y*, según la ecuación (2), donde $ID(I)$ indica la identidad de la persona representada por la imagen *I*.

$$y(I^a, I^b) = \begin{cases} 1 & ID(I^a) = ID(I^b) \\ 0 & ID(I^a) \neq ID(I^b) \end{cases} \quad (2)$$

La estimación de la matriz es un proceso iterativo, integrado en el proceso de aprendizaje de los descriptores que analiza. Por ello, en cada iteración, *t*, los datos no son un solo par, sino un conjunto (batch) de pares de descriptores, $P^t = \{F(I_i^a), F(I_i^b)\}$.

La estructura de los datos es representada por la distribución probabilística de la variable formada por la diferencia entre las características de los pares de imágenes $(F(I^a) - F(I^b))$. Dado que el método empleado es discriminativo, esta distribución es determinada de forma independiente para dos subespacios de características: el representado por los pares positivos, y el representado por los negativos. Por ello, dos conjuntos distintos son generados, P^+ y P^-, a partir de los pares positivos y negativos de *P*, como definen las ecuaciones (3) y (4), respectivamente, donde *B* es el número de pares de *P*.

$$P^+ = \left\{F(I_i^a) - F\left(I_i^b\right) \mid y\left(I_i^a, I_i^b\right) = 1,\ \forall i \in [1, B]\right\} \quad (3)$$

$$P^- = \left\{F(I_i^a) - F\left(I_i^b\right) \mid y\left(I_i^a, I_i^b\right) = 0,\ \forall i \in [1, B]\right\} \quad (4)$$

La distribución probabilística de cada subespacio de características es representada por su correspondiente matriz de covarianza, Σ^+y Σ^-. Sin embargo, si estas matrices son estimadas directamente a partir de los conjuntos P^+ y P^-, solo será considerada la información de los datos tratados en la presente iteración, y los valores de las matrices podrían oscilar fuertemente durante el proceso de entrenamiento.

Con el objetivo de considerar en cierta medida la contribución de los datos de iteraciones previas, desde una perspectiva estadística, y así obtener una muestra más representativa de los subespacios de características, se han empleado dos métodos diferentes para calcular Σ^+y Σ^-, descritos en las siguientes subsecciones. En primer lugar, se ha empleado el método basado en colas FIFO (First In, First Out), presentado en [5]. En segundo lugar, se ha diseñado un nuevo método basado en el algoritmo de la media móvil exponencialmente ponderada (EWMA, Exponentially Weighted Moving Average). En cada iteración, *t*, una vez que las matrices de covarianza han sido calculadas, la matriz de Mahalanobis es obtenida con la ecuación (5), cuya formulación fue presentada en [10].

$$M = (\Sigma_t^+)^{-1} - (\Sigma_t^-)^{-1} \quad (5)$$

2.2.1 Cálculo de las matrices de covarianza mediante colas FIFO.

En cada iteración, los elementos de P^+ y P^- son añadidos a dos colas de tamaño *K* llamadas Q^+ y Q^-, respectivamente. Cuando las colas están llenas, los elementos que fueron primeramente añadidos son eliminados para continuar añadiendo otros nuevos, ya que se trata de colas FIFO (First In, First Out), así, se almacenan las características aprendidas más recientemente. En cada iteración, *t*, las matrices de covarianza Σ^+ y Σ^-, son obtenidas a partir de los elementos almacenados en Q^+ y Q^-, como muestran las ecuaciones (6) y (7), donde $\mu_{t,j}^+$ y $\mu_{t,j}^-$, definidos por las ecuaciones (8) y (9), son el valor esperado para el elemento *j* del vector diferencia $(F(I^a) - F(I^b))$, para el subespacio de características positvo y negativo, respectivamente.

$$\Sigma_{t,fc}^+ = \frac{\sum_{i=1}^K \left(Q_{t,if}^+ - \mu_{t,f}^+\right)\left(Q_{t,ic}^+ - \mu_{t,c}^+\right)}{K} \quad (6)$$

$$\Sigma_{t,fc}^- = \frac{\sum_{i=1}^K \left(Q_{t,if}^- - \mu_{t,f}^-\right)\left(Q_{t,ic}^- - \mu_{t,c}^-\right)}{K} \quad (7)$$

$$\mu_{t,j}^+ = \frac{\sum_{i=1}^K Q_{t,ij}^+}{K} \quad (8)$$

$$\mu_{t,j}^- = \frac{\sum_{i=1}^K Q_{t,ij}^-}{K} \quad (9)$$

2.2.1 Cálculo de las matrices de covarianza mediante el método EWMA

Este método reduce el consumo de memoria computacional, ya que no requiere de colas para almacenar los datos de iteraciones anteriores. En su lugar, dos matrices de covarianza, $\Sigma_{P,fc}^+$ y $\Sigma_{P,fc}^-$, son

directamente calculadas a partir de los conjuntos P^+ y P^-, con las ecuaciones (10) y (11), respectivamente, por lo que tan solo representan la distribución para las características aprendidas en la iteración actual, *t*. El número de elementos en P es fijo (B=128), pero la proporción de pares positivos y negativos puede variar ligeramente en cada iteración, variando el tamaño de P^+ y P^-, representado por L^+ y L^-, respectivamente.

$\mu_{t,j}^+$ es el valor esperado para el elemento *j* del vector diferencia $(F(I^a) - F(I^b))$, para el subespacio de características positvo, y $\mu_{t,j}^-$, análogamente para el negativo. Para estimar estos valores, primero deben calcularse los valores esperados únicamente a partir de los elementos de P^+ y P^-, obteniendo $\mu_{P,j}^+$ y $\mu_{P,j}^+$. A continuación, $\mu_{t,j}^+$ y $\mu_{t,j}^+$, son obtenidos en cada iteración, *t*, con el método de la media móvil exponencialmente ponderada, EWMA, ecuaciones (14) y (15), teniendo en cuenta los valores obtenidos en la iteración anterior, *t*-1, y para las características aprendidas en la iteración actual, $\mu_{P,j}^+$ y $\mu_{P,j}^+$.

$$\Sigma_{P,fc}^+ = \frac{\sum_{i=1}^{L^+}\left(P_{t,if}^+ - \mu_{t,f}^+\right)\left(P_{t,ic}^+ - \mu_{t,c}^+\right)}{L^+} \quad (10)$$

$$\Sigma_{P,fc}^- = \frac{\sum_{i=1}^{L^-}\left(P_{t,if}^- - \mu_{t,f}^-\right)\left(P_{t,ic}^- - \mu_{t,c}^-\right)}{L^-} \quad (11)$$

$$\mu_{t,j}^+ = \beta\mu_{t-1,j}^+ + (1-\beta)\mu_{P,j}^+ \quad (12)$$

$$\mu_{t,j}^- = \beta\mu_{t-1,j}^- + (1-\beta)\mu_{P,j}^- \quad (13)$$

$$\mu_{P,j}^+ = \frac{\sum_{i=1}^{L^+} P_{t,ij}^+}{L^+} \quad (14)$$

$$\mu_{P,j}^- = \frac{\sum_{i=1}^{L^-} P_{t,ij}^-}{L^-} \quad (15)$$

Finalmente, para considerar la contribución de los datos analizados en iteraciones previas, no sólo en el cálculo de los valores esperados, sino también en el de las matrices de covarianza, $\Sigma_{t,fc}^+$ *y* $\Sigma_{t,fc}^-$, el método EWMA ha sido empleado de nuevo para estimar estas últimas, mediante las ecuaciones (16) y (17). Según estas ecuaciones, la matriz de covarianza para el subespacio de características positivo es calculada como una media ponderada entre la matriz obtenida en la iteración anterior, *t*-1, y la calculada a partir de las características aprendidas en la iteración actual. La matriz de covarianza para el subespacio de características negativo se estima de forma análoga.

$$\Sigma_{t,fc}^+ = \beta\Sigma_{t-1,fc}^+ + (1-\beta)\Sigma_{P,fc}^+ \quad (16)$$

$$\Sigma_{t,fc}^- = \beta\Sigma_{t-1,fc}^- + (1-\beta)\Sigma_{P,fc}^- \quad (17)$$

Tanto en el cálculo de los valores esperados como de las matrices de covarianza, el parámetro β es el peso dado a los valores obtenidos en la iteración previa, que fueron calculados a su vez, multiplicando β por los valores obtenidos previamente (*t*-2). Por lo tanto y dado que β es un número positivo y menor que uno, los pesos dados a las estimaciones obtenidas en cada iteración decrecen exponencialmente con un factor β, para iteraciones cada vez más tempranas. Para decidir el valor de β, se ha empleado la ecuación (18), que estima de manera aproximada, el alcance de la media, ρ, i.e. el número de iteraciones, cuyos valores obtenidos son considerados en la media.

$$\rho \approx \frac{1}{1-\beta} \quad (18)$$

3 PROCESO DE APRENDIZAJE

Tanto el entrenamiento de la red necesaria para extraer las características más discriminativas de una persona, como la estimación de la matriz de Mahalanobis son llevados a cabo simultáneamente en un mismo proceso de aprendizaje.

3.1 MODELO DE APRENDIZAJE SIAMÉS

Los descriptores y la matriz de Mahalanobis aprendidos deben ser tales que las distancias entre las imágenes de pares positivos sean menores que las de los pares negativos. Por lo tanto, la re-identificación puede ser resuelta como un problema de clasificación binaria (entre pares positivos y negativos), de manera que cada muestra a clasificar no es una imagen, sino un par de ellas. La figura 2 muestra ejemplos de pares positivos (uno en cada columna), formados por imágenes capturadas desde dos cámaras, *a* y *b*, pertenecientes a la base de datos PRID2011, [7], empleada para entrenar y evaluar el método propuesto.

Figura 2: Ejemplos de pares positivos de PRID2011

El aprendizaje se ha realizado mediante el modelo siamés. La red neuronal a entrenar, presentada en la sección 2.1, ha sido replicada en dos ramas, las cuales comparten los mismos parámetros, por lo que un único descriptor es aprendido, pero computado dos veces para la propagación de la red hacia delante (forward propagation) en su aprendizaje. Durante este paso, los descriptores de cada entrada de la red siamesa, es decir, las dos imágenes de un par, son calculados, $F(I^a)$ y $F(I^b)$, además de la distancia, *d*, entre ellos.

Finalmente, una función de coste, f_c, define el objetivo a alcanzar, es decir, el valor de distancia deseado, y mide la desviación de las distancias computadas durante el entrenamiento con respecto a tal objetivo. El valor de coste medido es usado en la propagación de la red hacia atrás (back-propagation), [17], para forzar los pesos de la red a obtener valores que hagan que el valor de la distancia se acerque a su objetivo. En este trabajo, se ha empleado la función de coste, presentada en [4], basada en la comparación de la distancia, previamente normalizada, *nd*, con los valores deseados, definidos por dos márgenes, m_1 y m_2, para pares positivos (y=1), y negativos (y=0). Esta función es definida por las ecuaciones (19) y (20), para un conjunto de *B* pares de entrenamiento. Según la ecuación (19), se considera que un par positivo está bien clasificado cuando la distancia, *nd*, entre sus imágenes es menor que m_1 ($m_1 = 0.3$), y uno negativo, cuando *nd* es mayor que m_2 ($m_2 = 0.7$).

$$f_c = \frac{1}{2B}\sum_{i=1}^{B}\begin{bmatrix} y_i \cdot max(nd_i - m_1, 0) + \\ (1 - y_i) \cdot max(m_2 - nd_i, 0) \end{bmatrix} \quad (19)$$

$$nd = 2\left(\frac{1}{1+e^{-d}} - 0.5\right) \quad (20)$$

3.2 INTEGRACIÓN DE LA DISTANCIA DE MAHALANOBIS EN EL APRENDIZAJE

En este trabajo se propone la distancia de Mahalanobis al cuadrado, d_M^2, ecuación (1), para comparar los descriptores aprendidos para dos imágenes.
Esta distancia requiere de la matriz de Mahalanobis para su cálculo, la cual es aprendida de forma simultánea a los descriptores. Por lo tanto, no se consigue una estimación fiable de esta matriz hasta alcanzar un cierto número de iteraciones, *T*. Por ello, en las primeras iteraciones del aprendizaje, se calcula la distancia entre los descriptores, *d*, con la distancia Euclídea al cuadrado, d_E^2, en lugar de la de Mahalanobis, como describe la ecuación (21).

$$d = \begin{cases} d_E^2(I^a, I^b) & t < T \\ d_M^2(I^a, I^b) & t \geq T \end{cases} \quad (21)$$

Se ha elegido un valor de *T* lo suficientemente grande para asegurar al menos la convergencia en el aprendizaje de las características, tras varias observaciones experimentales de la evolución de la función de coste durante el entrenamiento de la red.

4 RESULTADOS

En esta sección se describe la base de datos usada para entrenar y evaluar el método propuesto. Además, se analiza y se compara la evolución del proceso de aprendizaje para los dos métodos de estimación de las matrices de covarianza, presentados en la Sección 2.2, así como su capacidad de re-identificación.

4.1 DATASET

El algoritmo propuesto se ha evaluado en la base de datos PRID 2011 para re-identificación a partir de una sola captura (single-shot), siguiendo el protocolo estándar, definido en [7]. Se trata de una de las bases de datos más empleadas ya que se compone de imágenes capturadas desde dos vistas con importantes cambios de fondo, pose, iluminación y parámetros de las cámaras. La vista A contiene 385 imágenes, y la B, 749. Existen 200 individuos capturados desde ambas cámaras, de los cuales, 90 han sido seleccionados aleatoriamente como conjunto de entrenamiento, y 10 como conjunto de validación. Para realizar el test, las imágenes de la vista A correspondientes a las 100 personas restantes con representación en ambas vistas han sido utilizadas como conjunto de imágenes de prueba. La gallería de imágenes se ha formado con 649 imágenes, todas las de la vista B excepto las seleccionadas para el entrenamiento.

4.2 EXPERIMENTOS

Se han llevado a cabo dos entrenamientos diferentes. En ambos las características y la matriz de Mahalanobis han sido aprendidas mediante un modelo siamés, y se distinguen por el método empleado para calcular las matrices de covarianza (Sección 2.2): el método A, usando colas FIFO, con *K*=1000, y el método B a partir de la media exponencialmente ponderada, con $\beta = 0.94$. El número de parejas analizadas en cada iteración es de *B*=128. En [5], el valor elegido para K es 1000, de forma que las pilas almacenan los datos obtenidos en varias iteraciones, 16 iteraciones aproximadamente, ya que los conjuntos P^+ y P^- están formados por una media de B/2 parejas (1000/(128/2)=16). Esto es así, porque la proporción de parejas positivas y negativas formadas a partir de los individuos del conjunto de entrenamiento ha sido de 1:1. Para poder hacer una justa comparación de los métodos, se ha elegido un valor de β ($\beta = 0.94$) que haga que el alcance de las medias del método B sea de 16 iteraciones aproximadamente, ecuación (18).

Con el objetivo de comparar la evolución del proceso de entrenamiento para ambos experimentos, se han representado sus curvas de aprendizaje, esto es el valor de la función de coste en distintas iteraciones. Aunque la red se ha entrenado utilizando sólo el conjunto de entrenamiento, el coste se ha representado para ambos, el de entrenamiento y el de validación, en azul y naranja respectivamente, en la figura 3. Para ambas gráficas, el coste de entrenamiento rápidamente disminuye en las primeras iteraciones hasta hacerse prácticamente nulo. Sin embargo, crece drásticamente

en la iteración 50000, ya que este fue el valor dado a *T*, ecuación (21). A partir de esta iteración la comparación de las características es realizada con la distancia de Mahalanobis, en lugar de la Euclídea. Aunque el coste de entrenamiento es muy bajo en las primeras iteraciones no ocurre lo mismo con el de validación. Esto indica que la red está sobreentrenada, y no generaliza adecuadamente con ejemplos desconocidos. Sin embargo, en iteraciones posteriores el coste de validación disminuye gracias al aprendizaje de la matriz de Mahalanobis, que codifica las variaciones entre las dos vistas, reduciendo su efecto.

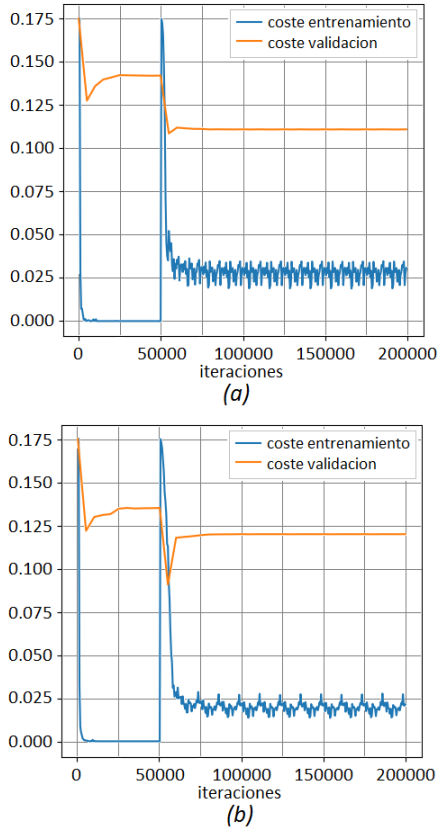

Figura 3: Curvas de aprendizaje para el método A (a) y el método B (b).

Aparentemente, el entrenamiento mediante el método B resulta más desfavorable, pues el coste de validación es ligeramente superior al ofrecido por el método A. Sin embargo, representando el porcentaje de parejas bien clasificadas, según la ecuación (19), para el grupo de entrenamiento y el de validación, puede observarse (Figura 4) que el método B presenta valores más altos, es decir, clasifica erróneamente un menor número de parejas, aunque el coste de las mismas fuese superior. Por lo tanto, el método B permite entrenar una red con mayor capacidad de discriminación entre pares positivos y negativos, resultando en una mejor re-identificación como se muestra en la siguiente sección.

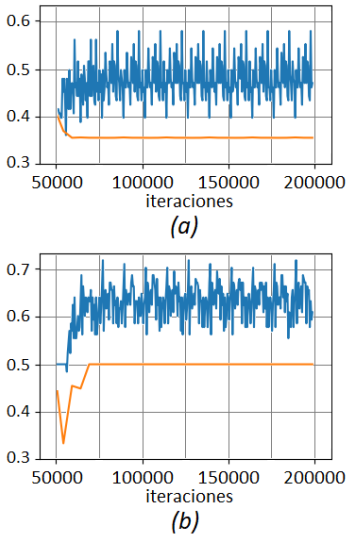

Figura 4. Porcentaje de parejas bien clasificadas, para el método A (a) y el método B (b).

4.3 EVALUACIÓN

Para evaluar la capacidad de re-identificación de la red entrenada con ambos métodos A y B (Sección 2.2), se ha calculado su curva acumulativa de porcentaje de aciertos (CMC, Cumulative Matching Characteristic), [13]. Para obtenerla, cada imagen de prueba es comparada con cada una de las de la galería, mediante el algoritmo propuesto. La curva representa, la probabilidad de encontrar el emparejamiento correcto (ambas imágenes representan la misma persona), tasa de acierto, entre los *r* mejores, i.e. los que presentan menores distancias. El valor *r* es llamado rango.

Los resultados obtenidos son mostrados por la tabla 1, dónde se han resaltado las tasas más altas de acierto para cada tango, *r*, y por la figura 5 para facilitar su comparación. Como se esperaba, la estimación de las matrices de covarianza mediante medias exponencialmente ponderadas, EMWA, proporciona un mejor aprendizaje tanto de las características como de la distancia que las relaciona, ya que, aunque considera datos de iteraciones pasadas, da mayor importancia a los más recientes, permitiendo una mejor actualización de la matriz de Mahalanobis y consecuentemente de los pesos aprendidos.

Las ventajas de realizar una estimación de la matriz de Mahalanobis simultánea al aprendizaje los descriptores de las imágenes son verificadas con los resultados mostrados en la tabla 2 y la figura 6. Se ha comparado el método propuesto (B), con otros métodos basados en la búsqueda de la combinación óptima de las características, como Rank-SVM, [15], y PRDC, [19], además de su fusión con un método para discriminar entre diferentes grupos de población, PSFI, [12]. Por otro lado, el uso del método de análisis

discriminativo propuesto muestra excelentes resultados en comparación con los ofrecidos por otros métodos discriminativos, LDA, [3], y LDML, [6].

Tabla 1: Tasas de acierto (en [%]) obtenidas con el método A y B.

Método	*r*=1	5	10	20	50	100
Método B	1	**15**	**21**	**28**	**38**	**58**
Método A	**2**	6	10	18	28	49

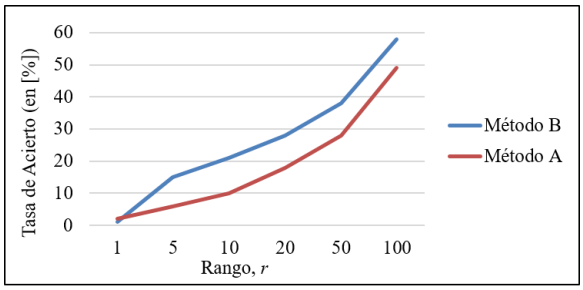

Figura 5. Curva CMC para el método A y B.

Tabla 2: Tasas de acierto (en [%]) para distintos algoritmos de re-identificación.

Método	*r*=1	10	20	50	100
Método b	1	**21**	**28**	**38**	**58**
LDA	**4**	14	21	35	48
PSFI+PRDC	3	9	16	24	39
PRDC	3	10	15	23	38
PSFI+RankSVM	**4**	9	13	20	32
RankSVM	**4**	9	13	19	32
LDML	2	6	11	19	32

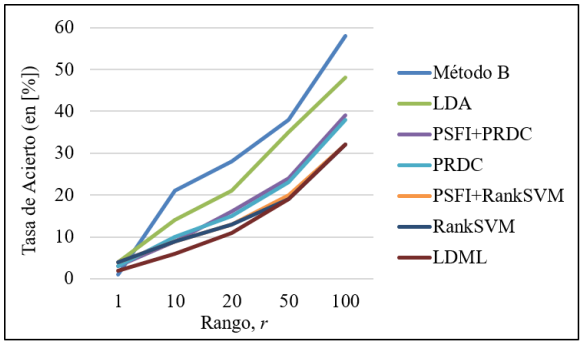

Figura 6. Comparación de curvas CMC para distintos algoritmos de re-identificación.

5 CONCLUSIONES

La re-identificación de una persona en dos imágenes se convierte en un reto desafiante cuando existen grandes variaciones de iluminación, perspectiva, escala o resolución entre las mimas. Este artículo propone un método para la codificación de tales variaciones en una matriz de Mahalanobis. El método presentado analiza la estructura de los datos en el espacio de las características de forma discriminativa, tratando la re-identificación como una clasificación binaria entre el subespacio positivo y negativo, que representan pares de imágenes correspondientes a una misma persona, o dos distintas, respectivamente. El análisis de los subespacios ha sido integrado en el proceso de aprendizaje de las características mediante el método de las medias móviles exponencialmente ponderadas (EWMA, Exponentially Weighted Moving Averages). De esta forma, el aprendizaje de la matriz de Mahalanobis y de los descriptores de las imágenes son efectuados simultáneamente, influyéndose y mejorándose recíprocamente. Todo ello es abordado mediante el entrenamiento de una red neural con una arquitectura siamesa, cuyo sobre-entrenamiento ha sido parcialmente apaliado por el uso de la distancia de Mahalanobis. Finalmente se ha obtenido un modelo capaz de generalizar su función discriminativa sobre ejemplos desconocidos, resultando en una mejora notable de la capacidad de re-identificación con respecto a otros métodos, como ha demostrado la evaluación realizada.

Agradecimientos

Este trabajo fue financiado por el Gobierno de España a través de dos proyectos CICYT (TRA2015-63708-R y TRA2016-78886-C3-1-R), la beca del Ministerio de Educación, Cultura y Deporte para la Formación de Profesorado Universitario (FPU14/02143), y la Comunidad de Madrid con SEGVAUTO-TRIES (S2013/MIT- 2713). Agradecemos la donación de la GPU utilizada en este trabajo a NVIDIA Corporation.

English summary

PERSON RE-IDENTIFICATION BY MAHALANOBIS DISTANCE

Abstract

Person re-identification requires the learning of a distance metric able to compare two images and decide if they belong, or not, to the same person. The automation of this task, in order to be applied in intelligent video-surveillance, involves a great challenge, due to the presence of people with similar appearance. For that reason, it is necessary to learn discriminative features and a metric to properly combine them. However, the variations of illumination, perspective, background, resolution and scale between two images of the same person, which were captured from different views, make his or her apperance vary, hampering the re-identification. This paper proposes coding the view-to-view tranformations in a Mahalanobis matrix, whose estimation has been integrated into the discriminative

features learning. In that way, these features can render the dissimilarity mainly due to appearance changes intead of the view changes. This estimation has been implemented as a new layer of a deep convolutional neural network, which has been trained and tested over the PRID2011 dataset.

Keywords: Mahalanobis Matrix, Deep Convolutional Neural Network, People Re-Identification.

Referencias

[1] Ahmed, E., Jones, M., Marks, T. K. (2015). "An improved deeplearning architecture for person re-identification", *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 3908-3916.

[2] Chan-Lang, S., Pham, Q. C., and Achard, C. (2016). "Bidirectional sparse representations for multi-shot person re-identification". *13th IEEE International Conference on Advanced Video and Signal Based Surveillance*, pp. 263–270.

[3] Fisher, R. A. (1936). "The use of multiple measurements in taxonomic problems", *Annals of eugenics*, 7(2), pp.179– 188.

[4] Gómez-Silva, M. J., Armingol, J. M., de la Escalera, A. (2017). "Deep part features learning by a normalised double-margin-based contrastive loss function for person re-identification", *Conference on Computer Graphics Theory and Applications*, pp. 277-285.

[5] Gómez-Silva, M. J., Armingol, J. M., de la Escalera, A. (2018). "Deep Part Similarity Learning for person re-identification", *Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, pp. 419 - 418.

[6] Guillaumin, M., Verbeek, J., and Schmid, C. (2009). "Is that you? metric learning approaches for face identification", *IEEE 12th international conference on Computer Vision*, pp. 498–505.

[7] Hirzer, M., Beleznai, C., Roth, P. M., and Bischof, H. (2011). "Person re-identification by descriptive and discriminative classification", *In Scandinavian conference on Image analysis*, pp. 91–102. Springer.

[8] Hirzer, M., Roth, P., K¨ostinger, M., and Bischof, H. (2012). "Relaxed pairwise learned metric for person reidentification", *Computer Vision–ECCV*, pp. 780–793.

[9] Hunter, J. S. (1986). "The exponentially weighted moving average", *Journal of quality technology*, 18(4), pp. 203-210.

[10] Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., and Bischof, H. (2012). "Large scale metric learning from equivalence constraints", *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp:2288–2295.

[11] Layne, R., Hospedales, T. M., and Gong, S. (2014). "Attributes-based re-identification", *Person Re- Identification*, pp. 93–117. Springer.

[12] Liu, C., Gong, S., Loy, C. C., and Lin, X. (2014). "Evaluating feature importance for re-identification". *In Person Re- Identification*, pp. 203–228. Springer.

[13] Moon, H., Phillips P. J., (2011) "Computational and performance aspects of pca-based face-recognition algorithms", *Perception*, 30(3): pp. 303–321.

[14] Munaro, M., Fossati, A., Basso, A., Menegatti, E., and Van Gool, L. (2014). "One-shot person reidentification with a consumer depth camera", *PersonRe-Identification*, pp. 161–181. Springer.

[15] Prosser, B., Zheng, W.-S., Gong, S., Xiang, T., and Mary, Q. (2010). "Person re-identification by support vector ranking", In BMVC, v.2, p. 6.

[16] Roth, P. M., Hirzer, M., Köstinger, M., Beleznai, C., and Bischof, H. (2014). "Mahalanobis distance learning for person re-identification", *In Person Re-Identification*, pp. 247–267. Springer.

[17] Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1988). "Learning representations by back-propagating errors", *Cognitive modeling*, 5(3):1.

[18] Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J. (2013). "Image classification with the fisher vector: Theory and practice", *International journal of computer vision*, 105(3), pp.222–245.

[19] Zheng, W.-S., Gong, S., and Xiang, T. (2011). "Person reidentification by probabilistic relative distance comparison", *IEEE conference on Computer vision and pattern recognition (CVPR)*, pp. 649–656.