Aplicación de la Kérnel Predictibilidad en el Registro y Segmentación de Imágenes.

Tesis que para obtener el grado de Doctor en Ciencias con Especialidad en Computación presenta:

Héctor Fernando Gómez García

Centro de Investigación en Matemáticas.

Guanajuato, Gto. Septiembre de 2008.

RESUMEN

En esta tesis, se explora una nueva medida de predictibilidad para variables aleatorias basada en el valor esperado de la evaluación de un kérnel adecuado sobre pares de muestras independientes, denominada kérnel predictibilidad. Basándose en esta medida, se generan varios algoritmos para el registro de imágenes, tanto paramétrico como no-paramétrico, los cuales presentan una gran robustez comparados con algorimos basados en información mutua. Se describe brevemente la utilización de esta medida como información a priori en el problema de segmentación de imágenes.

This thesis explores a new predictability measure for random variables, based on the expected value of kernel evaluations over independent pairs of samples. Several registration methods are derived from this measure and it is shown that their application offers a higher robustness in parametric and non-parametric problems, when compared to other methods based on mutual information. The use of this measure, as a priori information on image segmentation problems, is briefly described.

ÍNDICE GENERAL

1. INTRODUCCIÓN

El registro de imágenes representa un problema fundamental dentro del procesamiento digital de imágenes. Sus aplicaciones abarcan áreas muy diversas, entre las que sobresalen el procesamiento de imágenes médicas y la visión robótica. La entropía de Shannon ha sido utilizada tradicionalmente para evaluar la similitud entre imágenes provenientes de diferentes modalidades a través de la información mutua, sin embargo, dada la naturaleza del concepto de entropía (definido como el promedio de la información de una variable aleatoria), la evaluación de esta medida es altamente sensible a los rasgos poco importantes en las imágenes, lo cual reduce la robustez de estos métodos, sobre todo en problemas de registro en los cuales es necesario aplicar transformaciones espaciales de gran magnitud para alinear las imágenes. Para ejemplificar este punto, considérese el vector de probabilidades $\mathbf{p}_1 = [0.1\ 0.9]$, cuya entropía es de 0.47, si se actualiza este vector incrementando en 0.05 la primer entrada y decrementando la segunda en la misma magnitud, el nuevo vector, $\mathbf{p}_2 = [0.15\ 0.85]$, tendrá una entropía de 0.60. El incremento de la primer entrada aporta un 55 por ciento de la actualización de la entropía, a pesar de que esta entrada tiene una magnitud casi despreciable comparada con la segunda. Al registrar un par de imágenes, el valor de la entropía de la distribución conjunta se actualiza en cada iteración

del proceso de optimización; en una imagen, estas entradas están relacionadas con los rasgos menos importantes, por lo que la acción de la transformación espacial aplicada sobre estos rasgos se refleja fuertemente en la ganancia o pérdida de entropía conjunta.

En este trabajo se presenta una nueva medida para evaluar la predictibilidad de variables aleatorias, denominada kérnel predictibilidad. Basándose en la kérnel predictibilidad, se definen algunas medidas de similitud entre imágenes que presentan poca sensibilidad a los rasgos menos importantes en las imágenes (a diferencia de las medidas de similitud entre imágenes basadas en la entropía de Shannon), dando origen a diferentes estrategias de registro de imágenes de gran robustez.

Otro problema de gran importancia consiste en la segmentación de imágenes, a través de lo cual se busca asignar una etiqueta a cada punto de una imagen de manera que se identifiquen patrones regulares detro de ella, como pueden ser tonos de gris, texturas, o disparidades estereoscópicas. La aplicación de criterios sobre las propiedades del campo de etiquetas, mediante la construcción de una distribución a priori adecuada dentro de un enfoque bayesiano, es una técnica ampliamente utilizada y que puede producir resultados de gran calidad. En este trabajo se explora la aplicación de la kérnel predictibilidad en la segmentación de imágenes, definiendo una distribución a priori sobre el campo de etiquetas de gran generalidad. Esta nueva distribución a priori busca generar campos de etiquetas que presenten la mínima variabilidad dentro de grandes regiones homogéneas de la imagen.

Este documento, ha sido estructurado en siete capítulos. En el segundo de ellos se realiza una discusión de las principales medidas de dispersión e información, mientras que la medida de predictibilidad propuesta se describe en el tercero. El cuarto capítulo analiza con detalle el problema del registro de imágenes, con especial énfasis en la descripción de las metodologías basadas en medidas de información. La aplicación de la kérnel predictibilidad al problema de registro se detalla en el quinto capítulo. Finalmente, en el sexto capítulo se discute la aplicación de la kérnel predictibilidad en el problema de segmentación de imágenes a través de la definición de una nueva distribución a priori basada en la maximización de la kérnel predictibilidad de la distribución de etiquetas sobre regiones.

2. MEDIDAS DE DISPERSIÓN Y DE INFORMACIÓN

Resulta natural el intentar resumir las características de una variable aleatoria a través de diferentes medidas. Es interesante, por ejemplo, encontrar un valor numérico que sustituya adecuadamente a dicha variable en muchas aplicaciones; esta inquietud conduce al concepto de *medidas de tendencia central* y particularmente al de *valor esperado*. Una vez elegido un valor para sustituir a la variable aleatoria, puede ser necesario evaluar la calidad de la sustitución; una opción, consiste en medir la *dispersión* de la variable alrededor de su representación. Otro punto a considerar es la estimación de la dificultad con la que se pueden predecir los valores que adquirirá la variable antes de ser generados; lo que da origen a las llamadas *medidas de información*. Este capítulo está dedicado al análisis de estos puntos, realizando un enfoque especial en resaltar las ventajas y desventajas de las medidas de dispersión e información para representar las características de una variable aleatoria.

2.1. Valor Esperado y Otras Medidas de Tendencia Central.

El valor esperado, o promedio, de una variable aleatoria es sin duda la medida de tendencia central más común. Dada una variable aleatoria X, con distribución de probabilidad F, su valor esperado se define como:

$$E(X) = \int_X x dF(x) \ .$$

en donde la integral se extiende sobre todos los posibles valores de la variable X.

En muchas aplicaciones, sin embargo, la distribución de probabilidad que rige a la variable aleatoria es desconocida, por lo que debe realizarse una estimación del valor esperado basándose en un conjunto de muestreo formado por N variables independientes y distribuídas de acuerdo a F. Esta estimación se lleva a cabo a través de la *media aritmética*:

$$\mu_X = \frac{1}{N} \sum_i X_i \ .$$

en donde X_i y X_j son independientes para toda $i \neq j$ y $X_i \sim F \ , \forall i$.

El cálculo de la media aritmética es muy sensible a la presencia de datos atípicos, es decir, muestras que toman valores muy poco probables bajo F y muy diferentes al resto de los valores muestreados. En general, para evitar esta influencia, debe realizarse un análisis de los valores muestreados y descartar aquellos que claramente sean diferentes al resto. Lo anterior, sin embargo, implica establecer un criterio para clasificar una muestra como atípica, lo cual puede ser complicado.

El concepto de valor esperado se puede generalizar para considerar cualquier función, Φ, definida sobre los valores de X:

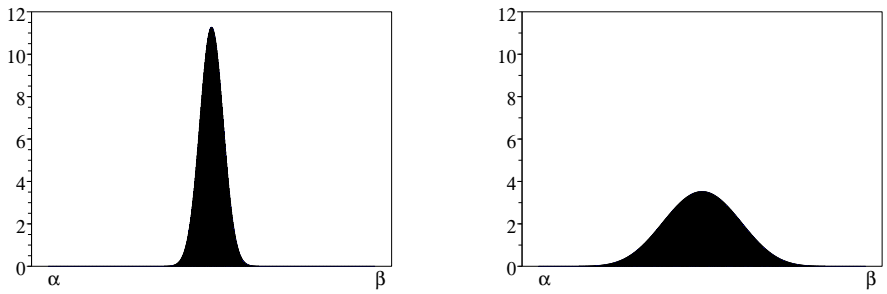
$$E[\Phi(X)] = \int_X \Phi(x) dF(x) \ .$$

La *mediana* es una medida de tendencia central más robusta ante datos atípicos que la media aritmética. La mediana se evalúa realizando una ordenación del conjunto de muestreo y eligiendo la muestra ubicada en la mitad del conjunto ordenado, cuando N es un número impar, o tomando el promedio de las dos muestras ubicadas en las posiciones $N/2$ y $N/2+1$ cuando N es par. Se debe notar que la robustez de la mediana ante datos atípicos se paga con un mayor costo computacional para su evaluación (de orden $N \log N$), comparado con la media aritmética, debido a la necesidad de ordenar las muestras.

Otras medidas de tendencia central, que presentan cierto grado de robustez ante datos atípicos son las siguientes:

- *media geométrica* $= \sqrt[N]{X_1 X_2 \cdots X_N}$
- *media armónica* $= \frac{N}{\sum_i \frac{1}{X_i}}$

Finalmente, dado un conjunto de muestras de una variable aleatoria discreta, se define la *moda*, como el valor que más veces se repite. En el caso de variables aleatorias ordenadas, es práctica común llamar moda a los valores en donde se ubican los máximos de la función de densidad.



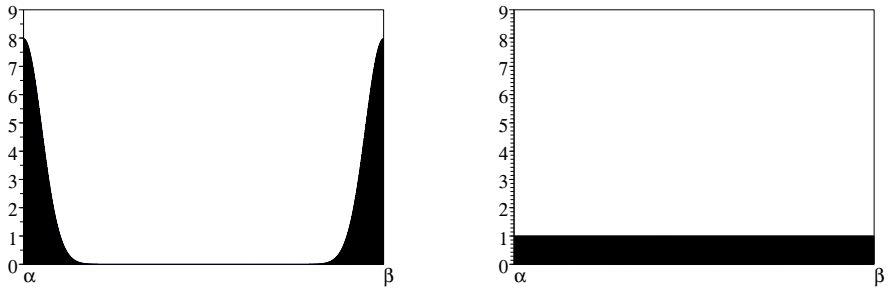
(a) (b)

Fig. 2.1: La variable aleatoria asociada con la densidad de probabilidad de la gráfica de la izquierda tiene una menor varianza (dispersión) que la de la derecha.

2.2. Medidas de Dispersión

La medida de dispersión más básica es el *rango*, el cual se define como la diferencia entre el máximo y el mínimo valor que una variable aleatoria puede obtener. Esta medida permite generar una idea acerca de los límites de una variable, sin embargo brinda poca información acerca del resto de sus valores. Una mejor opción consiste en medir la distancia de los valores de una variable con respecto a su media. La *varianza*, se define como el promedio de la diferencia al cuadrado entre los valores de la variable aleatoria, y su valor esperado:

$$Var(X) = E(X - E(X))^2 = \int_X [x - E(X)]^2 dF(x) \ .$$



(a) (b)

Fig. 2.2: La variable aleatoria asociada a la densidad de probabilidad de la gráfica izquierda tiene una mayor varianza que la asociada a la gráfica derecha, a pesar de que sus realizaciones se concentran en dos valores.

Variables cuyas realizaciones se concentran alrededor de la media tienen poca varianza, mientras que ésta aumenta a medida que la variable aleatoria adquiere valores alejados de la media con probabilidad no despreciable (ver figura 2.1). En el caso de distribuciones multimodales, sin embargo, la varianza podría reflejar la dispersión de los valores de una manera poco natural. Para ilustrar este punto, basta comparar la varianza de una variable bimodal, con modas equiprobables localizadas en los extremos de cierto intervalo $[a, b]$, con la de una variable con distribución uniforme sobre el mismo intervalo (ver figura 2.2). En el primer caso, la varianza tiende al valor $\frac{(b-a)^2}{4}$ mientras que en el segundo la varianza es tres veces menor, $\frac{(b-a)^2}{12}$, a pesar de que la primera distribución está concentrada sobre dos valores diferentes.

La *covarianza* entre dos variables aleatorias X e Y se define como:

$$Cov(X,Y) = E\left\{[X - E(X)][Y - E(Y)]\right\} \ .$$

La covarianza puede normalizarse dividiendo entre la raíz cuadrada del producto de las varianzas de cada variable, con lo que se obtiene el *coeficiente de correlación*:

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \ . \tag{2.1}$$

El valor de este coeficiente se encuentra limitado al intervalo $[-1,1]$. Al igual que la covarianza, los valores extremos del coeficiente de correlación se alcanzan cuando existe una dependencia lineal entre ambas variables y es igual a cero cuando son independientes.

2.3. Medidas de Información.

Para variables aleatorias multimodales, una medida que puede resultar más natural que la varianza, resulta de cuestionarse acerca de la dificultad para predecir el valor que tomará la variable antes de ser generado. Bajo este enfoque, los valores de la variable asociada a la distribución de la figura 2.2(b), deben ser más difíciles de predecir, a pesar de tener una varianza menor, que los de la variable asociada a la distribución de la figura 2.2(a). La evaluación de la predictibilidad de una variable aleatoria se realiza a través de las denominadas *medidas de información*. El concepto de *información* tiene su origen en el área de las comunicaciones [Sha48], en donde se busca transmitir y recuperar mensajes a través de algún canal de manera confiable. El envío de mensajes es equivalente al proceso de muestreo de una variable

aleatoria discreta X, con distribución de probabilidad $p(X)$. Los mensajes transmitidos pueden distorsionarse debido a la presencia de ruido en el canal, por lo que el receptor debe asegurarse de recuperar el valor originalmente transmitido. Este trabajo se facilita cuando X toma un pequeño conjunto de valores con mucha probabilidad, o lo que es lo mismo, cuando tiene pocas modas. La información se asocia directamente a la sorpresa generada por cada mensaje, ya que mensajes con mucha probabilidad de ser enviados conllevan poca sorpresa (información) al contrario de los mensajes poco probables. Agregando la condición de que la información trasmitida por dos fuentes independientes sea igual a la suma de la información de cada fuente, se llega a la siguiente definición para la información contenida en el mensaje x_i:

$$I(x_i) = \log_q \left[\frac{1}{p(X = x_i)} \right] . \tag{2.2}$$

En el contexto de comunicaciones la base del logaritmo, q, se selecciona de acuerdo a las características del canal de información (igual a dos en el caso de canales binarios, por ejemplo), aunque las propiedades de (2.2) no dependen de esta elección.

Se define la *entropía* de la variable aleatoria X como el valor esperado de la información:

$$H(X) = \sum_i p(X = x_i) \log_q \left[\frac{1}{p(X = x_i)} \right] . \tag{2.3}$$

Alternativamente, la entropía puede verse como una medida de la incertidumbre acerca de los valores que toma una variable aleatoria, siendo máxi-

ma al evaluarse sobre distribuciones uniformes y mínima sobre distribuciones aleatorias degeneradas (que puede tomar únicamente un valor fijo).

Si la variable aleatoria Y representa el valor del mensaje recibido en el otro extremo del canal de información, entonces puede obtenerse una medida que cuantifique la calidad del canal, realizando una comparación entre la incertidumbre acerca del mensaje enviado y la incertidumbre restante cuando se recibe el mensaje (cuando se conoce el valor de Y). Con ese objetivo, se define la *Información Mutua* entre las variables X e Y como la reducción en la entropía de X una vez conocido Y:

$$\begin{aligned} IM(X,Y) &= H(X) - H(X|Y) \\ &= H(X) + H(Y) - H(X,Y)\ . \end{aligned} \tag{2.4}$$

El concepto de entropía ha sido extendido y otras medidas de información han sido formuladas, por lo que se identifica a (2.3) como la entropía de Shannon. La entropía de Renyi, por ejemplo se define como:

$$R(X) = \frac{1}{1-\alpha} \ln \left(\sum_i p^\alpha (X = x_i) \right) , \tag{2.5}$$

y la entropía de Tsallis es igual a:

$$T(X) = \frac{1}{\alpha - 1} \left(1 - \sum_i p^\alpha (X = x_i) \right) , \tag{2.6}$$

en ambos casos α es un parámetro libre.

Cabe resaltar que haciendo $\alpha = 2$, en el caso de la entropía de Tsallis, se obtiene la conocida *entropía de Gini*, ampliamente utilizado en *machine learning*.

Dada una distribución de probabilidad discreta, $\mathbf{p} = \{p_1, p_2, \ldots, p_N\}$, su entropía es igual a la de cualquiera de las $N!$ distribuciones resultantes de aplicar una permutación a los valores del vector $\mathbf{p}$ (ver figura 2.3). Esta propiedad es adecuada para variables aleatorias categóricas, sin embargo, para variables aleatorias ordenadas, puede ser necesario la evaluación de la entropía y la varianza para reflejar con mayor precisión las características de la distribución. Supóngase, por ejemplo, que se clasifica la calidad de cierto producto con números enteros del 1 al 5 a través de algún criterio, asignando de forma ordenada los valores más bajos a los productos de menor calidad y los más altos a los mejores. Dos distribuciones de probabilidad se forman al considerar muestreos del producto en lotes diferentes. La primera distribución tiene una composición igual de muestras con la clasificación 1 y 5, mientras que la segunda está compuesta por muestras de calidad 4 y 5 con la misma proporción (ver imagen 2.4). Ambas distribuciones deben tener un valor de entropía igual a $\log(2)$, dada la propiedad de invariabilidad ante permutaciones de las entradas del vector de probabilidades, pero este resultado no refleja el hecho de que en el segundo caso la producción total, en general, puede clasificarse como de alta calidad. Una recategorización de los niveles, puede afectar drásticamente el valor de entropía, considérese por ejemplo combinar los niveles 4 y 5 en uno solo, la entropía de la segunda población para esta nueva categorización debería ser cero, mientras que la de

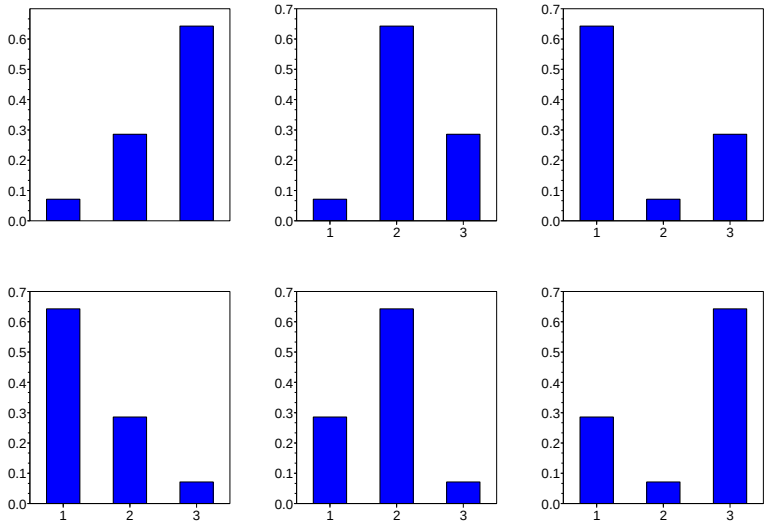
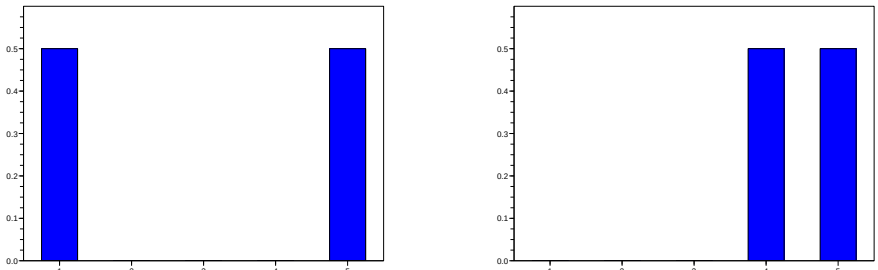


Fig. 2.3: Las distribuciones resultantes de permutar los valores de cualquier vector de probabilidades tienen la misma entropía.

la primer población no cambia.

La evaluación de la varianza de las dos distribuciones descritas anteriormente, brindaría una idea más clara acerca de sus características. Más aún, lo anterior también sugiere que una nueva medida de predictibilidad que considere la distancia entre los diferentes valores de la variable aleatoria, podría ser interesante.



(a) (b)

Fig. 2.4: Dos distribuciones de probabilidad con la misma entropía.

3. KERNEL-PREDICTIBILIDAD

Se puede introducir una medida de predictibilidad para una distribución de probabilidad F, considerando el siguiente juego aleatorio: alguien genera un valor x_1 de F y tratamos de adivinar el valor x_1, generando de forma independiente un nuevo valor x_2 de F. Otorgamos un premio a la predicción mediante una función $K(x_1, x_2)$. Al repetir este proceso se puede calcular el valor esperado de las evaluaciones de la función K para todos los pares de muestras generados. Si suponemos que la función K favorece predicciones cercanas al valor verdadero (K es una función decreciente de la distancia entre x_1 y x_2), entonces es claro que mientras menos incertidumbre contenga la distribución F más alto será el valor esperado del premio obtenido. Lo anterior conlleva a la siguiente definición para una distribución dada F:

$$KP(F) = E[K(X_1, X_2)] = \int_{R^d} \int_{R^d} K(x_1, x_2) dF(x_1) dF(x_2) \ . \tag{3.1}$$

Este funcional mide la predictibilidad de las variables aleatorias distribuidas de acuerdo a F, pesada por la función kérnel K, por lo cual se denominará *Kérnel-Predictibilidad* (KP). Debemos notar que KP es una medida de predictibilidad, a diferencia de la entropía, la cual es una medida de incertidumbre, por lo que ambas medidas tienen un comportamiento inverso al

evaluarse sobre diferentes distribuciones.

Considerando el caso de distribuciones discretas y ordenadas, el funcional (3.1) es equivalente a la siguiente expresión:

$$KP(\mathbf{p}) = \sum_i \sum_j K_{ij} p_i p_j = \mathbf{p}^T \mathbf{K} \mathbf{p} \tag{3.2}$$

donde la entrada (i, j) de la matriz $\mathbf{K}$ puede igualarse al premio otorgado por predecir el valor x_i cuando el valor generado es x_j, $K_{ij} = K(x_i, x_j)$. De manera general, este premio dependerá de la distancia entre x_i y x_j, por lo que la matriz $\mathbf{K}$ puede suponerse simétrica, sin embargo esto no es una restricción.

Al tomar en cuenta la distancia entre los valores de la variable aleatoria, la kérnel-predictibilidad de un vector de probabilidad no es necesariamente invariante ante permutaciones de sus entradas, a diferencia de la entropía. Por ejemplo, haciendo $K_{i,j} = e^{-|i-j|}$, en la matriz de (3.2), la kérnel-predictibilidad de la distribución mostrada en la figura 2.4(a) es igual a 0.509, mientras que la de la distribución de la figura 2.4(b) es igual a 0.684.

Es posible evaluar analíticamente la kérnel-predictibilidad para algunas distribuciones de probabilidad y kérneles específicos. Por ejemplo, si X es una variable con distribución Bernoulli, y $p = P(X = 1)$, entonces:

$$KP(X) = \left[(1-p)^2 + p^2\right] k_0 + 2p(1-p)k_1 \ , \tag{3.3}$$

siendo $k_0 = K_{i,j}$ si $i = j$ y $k_1 = K_{i,j}$ si $i \neq j$. Esta función, de p, tiene tres

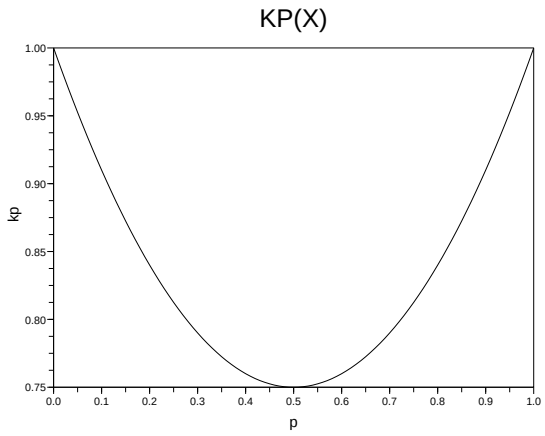


Fig. 3.1: Gráfica de kérnel-predictibilidad para una variable, X, con distribución Bernoulli y $p(X = 1) = p$.

extremos, un mínimo en $p = \frac{1}{2}$, en donde $KP(X) = \frac{1}{2}(k_0 + k_1)$, y alcanza su valor máximo, igual a k_0 (suponiendo que $k_0 > k_1$), para $p = 0$ y $p = 1$, lo cual coincide con la intuición. La figura 3, muestra la gráfica de la función (3.3) para $k_0 = 1$ y $k_1 = 0.5$.

Entre las opciones posibles para la función K, una muy natural es el kérnel gaussiano, el cual está definido en la siguiente expresión:

$$K(x_1, x_2) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right) \tag{3.4}$$

donde σ es un parámetro que permite controlar la amplitud del kérnel, d es la dimensión de la distribución multivariada y $\|\cdot\|$ es la distancia euclidiana en R^d.

Si se utiliza un kérnel gaussiano para medir la kérnel-predictibilidad de una variable aleatoria $d-dimensional$, X, con distribución gaussiana se obtiene lo siguiente:

$$\begin{aligned} KP(X) &= k_1 k_2^2 \int_{X_1}\int_{X2} e^{-\left[\frac{\|x_1-x_2\|^2}{2\sigma_1^2}+\frac{\|x_1-\mu\|^2}{2\sigma_2^2}+\frac{\|x_2-\mu\|^2}{2\sigma_2^2}\right]} dx_2 dx_1 \\ &= k_2 \int_{X_1} e^{-\frac{\|x_1-\mu\|^2}{2\sigma_2^2}} \left[k_1 k_2 \int_{X_2} e^{-\frac{\|x_1-x_2\|^2}{2\sigma_1^2}} e^{-\frac{\|\mu-x_2\|^2}{2\sigma_2^2}} dx_2\right] dx_1 \\ &= \frac{k_3}{\left[2\pi(\sigma_1^2+2\sigma_2^2)\right]^{d/2}} \int_{X_1} e^{-\frac{\|x_1-\mu\|^2}{2\sigma_3^2}} dx_1 \\ &= \frac{1}{\left[2\pi(\sigma_1^2+2\sigma_2^2)\right]^{d/2}} \end{aligned} \tag{3.5}$$

en donde se ha aplicado la propiedad de convolución de dos gaussianas; y además, σ_1 es el parámetro de amplitud del kérnel, σ_2 el de la distribución gaussiana, $\sigma_3^2 = \frac{\sigma_2^2(\sigma_1^2+\sigma_2^2)}{(\sigma_1^2+2\sigma_2^2)}$ y $k_i = \frac{1}{[2\pi\sigma_i^2]^{d/2}}$, $i \in \{1,2,3\}$.

Como era de esperarse, el resultado resumido en (3.5) muestra que la kérnel predictibilidad de la distribución gaussiana es inversamente proporcional a su varianza, obteniendo el máximo valor cuando $\sigma_2 \to 0$, y el mínimo cuando $\sigma_2 \to \infty$.

Algunas medidas que pueden ser confundidas con KP han sido presentadas con anterioridad, sin embargo una diferencia importante debe remarcarse. En [ZC04] un funcional similar a (3.1) se utiliza para calcular el valor esperado de la distancia entre dos grupos de imágenes. Mientras que [YDD05] y [SAA04],

presentan medidas de similitud entre imágenes que pueden confundirse con uno de los estimadores para (3.1) (el cual se discutirá más adelante). Sin embargo, estas tres medidas se evalúan sobre dos distribuciones diferentes, en contraste a (3.1), la cual toma una sola distribución en su argumento y representa una propiedad de la distribución tal como su entropía o su varianza.

3.1. Kérnel-Predictibilidad con Kérneles Gaussianos y Ventanas de Parzen Gaussianas.

Es posible extender aún más el resultado presentado en (3.5), si se realiza una aproximación no paramétrica, f_X, de la densidad de una distribución arbitraria F, mediante ventanas de Parzen gaussianas [DH73] centradas sobre un conjunto de muestras independientes $\{a_i\}$ obtenidas de F:

$$f_X(x) = \frac{1}{N}\sum_{i=1}^{N} f_{a_i,\sigma_2}(x) \tag{3.6}$$

en donde $f_{a_i,\sigma_2}(x)$ es la densidad gaussiana multivariada, $\mathcal{N}(a_i, \sigma_2^2\mathbf{I})$, con matriz de covarianza $\sigma_2^2\mathbf{I}$ de orden $d \times d$. Al emplear un kérnel gaussiano multivariado para medir $KP(F)$, con parámetro de amplitud σ_1 y utilizando de nueva cuenta el hecho de que la convolución de dos gaussianas es otra gaussiana, puede mostrarse que:

$$KP(F) = \frac{1}{N\left[2\pi(\sigma_1^2 + 2\sigma_2^2)\right]^{d/2}} \sum_i \sum_j \exp(-||a_i - a_j||^2/2(\sigma_1^2 + 2\sigma_2^2)). \tag{3.7}$$

Nótese que mientras mayor sea la separación de los puntos $\{a_i\}$ muestreados de la distribución F, menor será el valor de KP correspondiente. El máximo se obtiene cuando todos los puntos en el conjunto $\{a_i\}$ son iguales, lo cual representa una distribución definida por una sola gaussiana. En este caso, el valor de KP varía inversamente con respecto a σ_2, lo que implica que el máximo valor de KP se obtiene cuando la varianza de la distribución es igual a cero, o de manera equivalente cuando F sea la distribución de una variable aleatoria degenerada. Lo anterior es también válido para el caso discreto y para kérneles arbitrarios, siempre y cuando los elementos en la diagonal principal de la matriz $\mathbf{K}$ contengan el máximo valor K_M (recompensa máxima otorgada por una predicción exacta). Esto se deriva de la siguiente desigualdad:

$$KP(\mathbf{p}) = \sum_i \sum_j K_{ij} p_i p_j \leq K_M \sum_i \sum_j p_i p_j = K_M$$

y del hecho de que K_M es el valor de KP obtenido para distribuciones de variables aleatorias degeneradas.

En el caso de distribuciones de probabilidad continuas, es también posible mostrar el efecto de que la KP es sensible a la permutación de sus valores a diferencia de la entropía, lo cual se deduce de la ecuación (3.7) y se ilustra en la figura 3.1. Si la ventana gaussiana centrada sobre el punto a_1 se traslada hacia a_1^*, lo cual equivale a mover una porción de la masa de la distribución a una posición donde prácticamente no existe traslape con la distribución original, el valor de KP se reducirá, dado que la separación de los puntos del conjunto $\{a_i\}$ aumenta; al mismo tiempo la entropía de la distribución se incrementa. No obstante, si a_1 se traslada hasta un punto a_1^{**} situado todavía

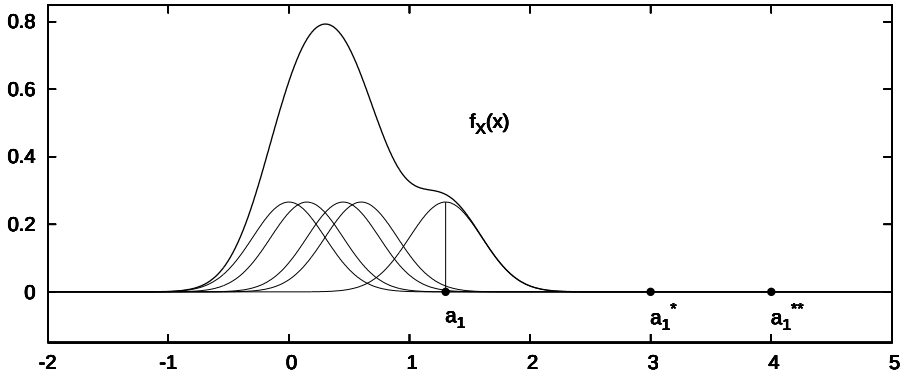


Fig. 3.2: Mover la ventana gaussiana centrada en a_1 hacia a_1^* reduce la entropía y la KP. La KP seguirá reduciéndose al mover a_1 aún más a la derecha, mientras que la entropía permanece constante.

más a la derecha, el valor de KP se reducirá aún más, sin embargo la entropía permanecerá prácticamente constante. Esta propiedad de la entropía es una desventaja cuando se aplica en problemas como el registro de imágenes, en donde la calidad de una transformación espacial se mide por la concentración de la distribución de tonos de gris conjunta entre un par de imágenes; en este caso el gradiente de KP contendrá más información sobre la localización de la transformación óptima.

La construcción de la matriz $\mathbf{K}$ en (3.2) mediante kérneles gaussianos, produce dos casos interesantes al evaluar el kérnel en valores extremos del parámetro de amplitud σ. En el primer caso, que corresponde a valores de σ muy pequeños, el kérnel gaussiano puede aproximarse por la delta de Kronecker de la siguiente manera:

$$G(x_i, x_j) = \delta_{ij} = \begin{cases} 1 & \text{if } x_i = x_j \\ 0 & \text{en otro caso.} \end{cases} \tag{3.8}$$

y entonces $KP(\mathbf{p}) = 1 - Gini(\mathbf{p})$; en donde *Gini* es la *entropía de Gini*, definida sobre la entropía de Tsallis (2.6). La entropía de Gini es maximizada, y el correspondiente valor de KP minimizado, al evaluarse con la distribución uniforme.

El segundo caso se deriva al utilizar valores de σ grandes, dado que empleando una aproximación en serie de Taylor, el kérnel gaussiano puede escribirse como:

$$G_\sigma(x_1, x_2) \approx 1 - \frac{\|x_1 - x_2\|^2}{2\sigma^2} \tag{3.9}$$

y para este caso, $KP(\mathbf{p}) \approx 1 - \frac{\sum_i Var[(X)_i]}{\sigma^2}$; en donde $Var[(X)_i]$ es la varianza del *i-ésimo* elemento de la variable aleatoria multivariada X. Como se mostró anteriormente (ver figura 2.2), para distribuciones univariadas con dominio finito sobre algún intervalo $[a, b]$, aquellas cuya densidad tiende a concentrarse simétricamente en sus dos valores extremos, a y b, tienen mayor varianza, y por lo tanto menor KP, que la distribución uniforme sobre el mismo dominio.

Tomando en cuenta que variables aleatorias con distribución uniforme son más difíciles de predecir que variables que toman únicamente dos valores con la misma probabilidad, es deseable que KP tenga un comportamiento similar a la entropía de Gini, y por esta razón se selecciona un valor pequeño para el

parámetro de amplitud del kérnel gaussiano; en la práctica, se toma σ entre el 2 % y el 10 % del rango de la variable aleatoria.

3.2. Estimación de la Kérnel-Predictibilidad

La expresión (3.1) representa un *funcional estadístico regular* de grado dos (dos es el número de argumentos del kérnel), y para su estimación tres diferentes opciones se encuentran regularmente en la literatura [Lee90][Leh99]. Estos estimadores se basan siempre en la utilización de un conjunto de muestreo compuesto por n variables aleatorias independientes e idénticamente distribuídas, $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, con $X_i \sim F$, $\forall i$; y se definen como:

$$\widehat{KP}^1 = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} K(X_i, X_j) \tag{3.10}$$

$$\widehat{KP}^2 = \frac{4}{n^2} \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} K(X_i, X_j) \tag{3.11}$$

$$\widehat{KP}^3 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K(X_i, X_j)\ . \tag{3.12}$$

Considerando que $KP(F) = E\left[K(X_i, X_j)\right]$ para $i \neq j$, puede verse que los estimadores (3.10) y (3.11) son insesgados, mientras que el estimador (3.12) tiene un sesgo inversamente proporcional al tamaño del conjunto de muestreo, lo cual se deduce en la siguiente expresión:

$$
\begin{aligned}
E\left(\widehat{KP}^3\right) &= E\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K(X_i,X_j)\right] \\
E\left(\widehat{KP}^3\right) &= E\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}K(X_i,X_j)+\frac{1}{n^2}\sum_{i=1}^{n}K(X_i,X_i)\right] \\
E\left(\widehat{KP}^3\right) &= \frac{n-1}{n}KP(F)+\frac{1}{n}E\left[K(X_1,X_1)\right] \\
E\left(\widehat{KP}^3\right) &= KP(F)+\frac{1}{n}\left\{E\left[K(X_1,X_1)\right]-KP(F)\right\} \ .
\end{aligned}
$$

Si el kérnel es simétrico, entonces $\widehat{KP}^1$ es el estimador de mínima varianza entre todos los estimadores insesgados, tal como se demuestra en [Lee90][Leh99]. El estimador $\widehat{KP}^2$ tiene más varianza que $\widehat{KP}^1$ sin embargo el costo computacional de su evaluación es el menor. Finalmente, el estimador $\widehat{KP}^3$ tiene la menor varianza entre estos tres estimadores. Debe notarse que al incrementar el tamaño del conjunto de muestreo, las varianzas de los tres estimadores disminuyen y tienden al mismo valor, mientras que el sesgo del tercer estimador tiende a cero.

3.3. Incremento de Kérnel-Predictibilidad

Considerando la versión discreta de la KP (ecuación 3.2) como una función de las entradas del vector de probabilidades, y realizando una expansión en serie de Taylor, los incrementos en KP, que podrían asociarse a algún proceso de optimización, son determinados por la siguiente expresión:

$$\Delta KP = 2\sum_i\left(\sum_j K_{ij}p_j\right)\Delta p_i \ .$$

Nótese que el incremento de cada elemento del vector de probabilidades, p_i, se multiplica por el coeficiente $\left(\sum_j K_{ij}p_j\right)$; el cual es equivalente al *i-ésimo* elemento del vector generado por el producto de la matriz $\mathbf{K}$ con el vector de probabilidades $\mathbf{p}$. Este producto equivale a un suavizamiento del vector $\mathbf{p}$ si asumimos que los valores K_{ij} son mayores mientras más cercanos estén de la diagonal principal. Por consecuencia, $\left(\sum_j K_{ij}p_j\right)$ es mayor para los valores de p_i más grandes, y el incremento en KP está determinado por las entradas del vector de probabilidad con mayor magnitud. Lo anterior representa una diferencia importante con respecto a la entropía como se verá más adelante.

4. REGISTRO DE IMÁGENES

El problema de registro de imágenes ha sido ampliamente explorado debido a la importancia de sus aplicaciones (ver [Got92][MV98][JPMV03][ZF03] y referencias ahí contenidas), las cuales abarcan áreas tales como el análisis de imágenes médicas, la visión robótica, la realidad aumentada, entre otras. Dadas dos imágenes I_S e I_R, las cuales pueden identificarse como *imagen fuente* e *imagen de referencia* respectivamente, registrar ambas imágenes es equivalente a encontrar una transformación espacial general T que una vez aplicada a I_S, permita alinear las estructuras comunes en I_R e I_S (ver fig. 4). Durante las últimas décadas una gran cantidad de métodos para registrar imágenes han sido presentados. Muchos de ellos se basan en la optimización de alguna medida de similitud o de diferencia entre la imagen transformada, $I_T = I_S(T)$, y la imagen de referencia I_R. La medida de similitud debe cuantificar la calidad de la transformación T para alinear I_S e I_R, lo que permite replantear el registro como un problema de optimización sobre el espacio de las transformaciones espaciales. La selección adecuada del método de registro se realiza considerando una serie de factores como lo son el tipo de transformación a aplicar y la relación entre las intensidades de las imágenes.

En la selección de la transformación a aplicar deben tomarse en cuenta las causas del desalineamiento entre I_S e I_R. En algunos casos puede ser

Fig. 4.1: Imagen fuente (izquierda), imagen transformada (centro) e imagen objetivo (derecha).

suficiente el aplicar una transformación con pocos grados de libertad, por ejemplo al registrar imágenes de una misma escena obtenidas desde diferentes puntos de vista. En esos casos el registro puede realizarse mediante transformaciones rígidas, afines o polinomiales en general, las cuales, al estar determinadas por un número reducido de valores se denominan *transformaciones paramétricas*. Sin embargo, en algunos problemas puede ser necesario incrementar el número de grados de libertad de la transformación para alinear adecuadamente las imágenes, por ejemplo al registrar imágenes médicas de diferentes personas [HBC+03], llegando al extremo de aplicar un vector de traslación diferente en cada punto. Bajo estas circunstancias, el número de incógnitas que debe determinarse está directamente relacionado con las dimensiones de las imágenes a registrar. Estas tranformaciones se denominan *transformaciones no-paramétricas*. El registro de imágenes utilizando transformaciones no-paramétricas es un problema mal condicionado debido a la existencia de múltiples soluciones en regiones sin textura y debe incorporar

restricciones sobre el espacio de soluciones, que por lo general se realiza imponiendo condiciones de regularidad sobre el campo vectorial que determina la transformación [HR81].

Otro punto a considerar es la relación entre los valores de intensidad de las imágenes. El problema de registro se facilita cuando se conoce la manera en que se transforma la intensidad de un punto en una imagen, para generar el valor de intensidad del punto correspondiente en la otra. En muchos casos el registro puede realizarse suponiendo que los valores de intensidad permanencen constantes entre las imágenes, sin embargo, esta suposición puede ser violada fácilmente, por ejemplo, cuando se presentan cambios de iluminación entre las imágenes, o al trabajar con imágenes médicas obtenidas mediante fuentes diferentes (resonancias magnéticas y tomografías computarizadas por ejemplo). Para enfrentar este problema, algunos métodos de registro buscan encontrar una función de transferencia de tonos que modele los cambios de intensidad entre puntos correspondientes al mismo tiempo que determinan la transformación geométrica que los alinea [Neg98, TLCH02, KK06]; sin embargo, la aplicación exitosa de estos métodos es limitada ya que algunos problemas presentan cambios de intensidad entre las imágenes que son imposibles de explicar mediante una simple función, sobre todo cuando los cambios de intensidad dependen también de la ubicación espacial de los puntos. Para este tipo de problemas, el registro de imágenes mediante la maximización de la información mutua (2.4) ha sido ampliamente utilizado desde su introducción a mediados de la década pasada, ya que para su aplicación no es necesario tener conocimiento de la forma en que se relacionan los valores de

intensidad entre las puntos correspondientes de ambas imágenes, sino que se basa en la idea de que al alinearlas se puede obtener la máxima dependencia (información) entre sus intensidades.

En las siguientes secciones se realiza una descripción de las principales metodologías aplicadas tradicionalmente en la solución del problema de registro.

4.1. Métodos Basados en la Restricción de Flujo Óptico.

Las imágenes I_S e I_R pueden representar un par de muestras temporales de una función, $I(\mathbf{x}, t)$, que define la intensidad de cualquier punto $\mathbf{x}$ del plano de una imagen a través del tiempo. Si se supone que los cambios temporales en el plano de la imagen únicamente redistribuyen espacialmente el valor de intensidad de cada punto, entonces la evolución de I se explica mediante el movimiento de pequeñas partículas ubicadas en la posición $\mathbf{x}$, en el instante t, con intensidad constante igual a $I(\mathbf{x}, t)$. La posición de cada partícula es también una función del tiempo con lo que se escribe $I[\mathbf{x}(t), t]$. Por lo anterior, la dinámica de la función I se modela mediante una ley de conservación, en la cual la integral $\int_\Omega I\left[\mathbf{x}(t), t\right] d\mathbf{x} = C$, para cualquier valor de t (siendo Ω el plano de la imagen y C una constante). Al derivar la integral anterior con respecto al tiempo, se producen las siguientes igualdades:

$$\frac{d}{dt}\int_{\Omega} I\left[\mathbf{x}(t),t\right]d\mathbf{x} = 0$$

$$\int_{\Omega} \frac{d\left\{I\left[\mathbf{x}(t),t\right]\right\}}{dt}d\mathbf{x} = 0$$

$$\int_{\Omega}\left[\frac{\partial I}{\partial t} + \nabla_{\mathbf{x}} I^{T}\frac{d\mathbf{x}}{dt}\right]d\mathbf{x} = 0\,. \tag{4.1}$$

El vector $\frac{d\mathbf{x}}{dt}$ define la velocidad de la partícula ubicada en la posición $\mathbf{x}$.

Nótese que para que la ecuación (4.1) se cumpla, es necesario que:

$$\frac{\partial I}{\partial t} + \nabla_{\mathbf{x}} I^{T}\frac{d\mathbf{x}}{dt} = 0 \tag{4.2}$$

para todos los valores de $\mathbf{x}$ y t. Esta ecuación diferencial es conocida como la *restricción de flujo óptico*.

El campo vectorial que se genera al considerar los vectores de velocidad de todas las partículas en un instante t_1, permite realizar el alineamiento de los rasgos de la imagen obtenida en t_1 con los rasgos de la imagen para un instante posterior t_2 (siempre y cuando t_1 y t_2 sean valores muy cercanos); por lo que el registro de las imágenes se realiza encontrando $\frac{d\mathbf{x}}{dt}$ para toda $\mathbf{x}$. Bajo estas condiciones, el registro de imágenes puede explicar el movimiento de las componentes de una escena dinámica.

La ecuación diferencial (4.2), tiene soluciones múltiples en regiones en las que $\nabla_{\mathbf{x}} I = 0$ (regiones homogéneas), por lo que es necesario imponer más

restricciones. En la práctica, para realizar el registro, se busca el campo vectorial que satisfaga:

$$T = \min_{\mathbf{u}} \left\{ \int_{\Omega} \left\| \frac{\partial I}{\partial t} + \nabla_{\mathbf{x}} I^T \mathbf{u}(\mathbf{x}) \right\|^2 d\mathbf{x} + \lambda \int_{\Omega} V\left[\mathbf{u}(\mathbf{x})\right] d\mathbf{x} \right\} \tag{4.3}$$

Originalmente [HR81], la función V se seleccionó como la suma del cuadrado de la magnitud de los gradientes de cada componente del campo vectorial, $V[\mathbf{u}(\mathbf{x})] = \sum_i \|\nabla u_i(\mathbf{x})\|^2$, o como la suma del cuadrado de los laplacianos, $V[\mathbf{u}(\mathbf{x})] = \sum_i [\Delta u_i(\mathbf{x})]^2$. Sin embargo, diversas modificaciones han sido propuestas al funcional (4.3), ya que la utilización de potenciales cuadráticos en los dos términos genera soluciones que no permiten capturar discontinuidades del flujo óptico, además de ser muy sensibles a la presencia de datos atípicos. El uso de potenciales robustos ha sido explorado en [BR96, ADK99], mientras que en [BHS97, KK06] la restricción de flujo óptico se integra mediante el concepto de mínima mediana de cuadrados (Least Median Squares) en contraste con el enfoque de mínimos cuadrados originalmente propuesto.

Como se mencionó anteriormente, la conservación de la intensidad, establecida en la restricción de flujo óptico, es una condición que difícilmente se cumple. En la práctica, simples cambios de iluminación en la escena pueden comprometer la aplicación exitosa de estos métodos. Aunque la restricción de flujo óptico puede ampliarse para modelar dichos cambios [Neg98, TLCH02, KK06], los modelos actuales distan mucho de ofrecer una buena solución en problemas realistas. Más aún, la restricción de flujo óptico puede ser inaplicable en problemas en los que las imágenes provienen de diferentes sensores, y en donde el registro es necesario para integrar información complementaria.

4.2. Registro de Imágenes por Métodos Espectrales.

Supóngase que la imagen I_R se obtiene aplicando un vector de traslación constante a la imagen I_S:

$$I_R(x) = I_S(x - x_0) \ ,$$

entonces, aplicando la propiedad de traslación, las transformadas de Fourier de ambas imágenes están relacionadas por la siguiente ecuación:

$$\widehat{I}_R(\omega) = e^{-i\omega^T x_0}\widehat{I}_S(\omega) \ , \tag{4.4}$$

en donde $\widehat{I}$ es la transformada de Fourier de I.

Utilizando la representación polar, la anterior ecuación se reescribe como:

$$\begin{aligned} |\widehat{I}_R(\omega)|e^{i\theta_R} &= e^{-i\omega^T x_0}|\widehat{I}_S(\omega)|e^{i\theta_S} \\ |\widehat{I}_R(\omega)|e^{i\theta_R} &= |\widehat{I}_S(\omega)|e^{i(\theta_S - \omega^T x_0)} \ , \end{aligned}$$

de donde se desprende que $|\widehat{I}_R(\omega)| = |\widehat{I}_S(\omega)|$ y que:

$$e^{i(\theta_R - \theta_S)} = e^{-i(\omega^T x_0)} \ , \tag{4.5}$$

por lo que el efecto de haber trasladado espacialmente la imagen I_S equivale a aplicar una diferencia de fase, de magnitud $\omega^T x_0$, sobre su transformada de Fourier.

Tomando en cuenta que la fase de un número complejo se obtiene dividiendo dicho número por su magnitud, la ecuación (4.5) puede reescribirse como:

$$\frac{\widehat{I}_R(\omega)}{|\widehat{I}_R(\omega)|}\frac{\widehat{I}_S^*(\omega)}{|\widehat{I}_S(\omega)|} = e^{-i(\omega^T x_0)} \ , \tag{4.6}$$

en donde $\widehat{I}_S^*(\omega)$ es igual al complejo conjugado de $\widehat{I}_S(\omega)$. El término del lado izquierdo de la ecuación anterior es conocido como *correlación de fase*.

Al calcular la transformada de Fourier inversa de (4.6) se obtiene:

$$F^{-1}\left[\frac{\widehat{I}_R(\omega)}{|\widehat{I}_R(\omega)|}\frac{\widehat{I}_S^*(\omega)}{|\widehat{I}_S(\omega)|}\right](x) = \delta(x - x_0) \ . \tag{4.7}$$

Y de lo anterior, el valor del vector de traslación aplicado a I_S puede recuperarse encontrando:

$$x_0^* = \arg\max_x F^{-1}\left[\frac{\widehat{I}_R(\omega)}{|\widehat{I}_R(\omega)|}\frac{\widehat{I}_S^*(\omega)}{|\widehat{I}_S(\omega)|}\right](x) \ . \tag{4.8}$$

Algunos métodos de registro de imágenes basados en la correlación de fase pueden encontrarse en las siguientes referencias [Hog03, KAM04, WCLY06]. Cabe hacer notar que el valor de x_0 estimado en la expresión (4.8), está limitado a contener componentes enteras, pues el dominio de la imagen es discreto. Existen diferentes estrategias para poder recuperar traslaciones a nivel de subpixel, una de ellas consiste en encontrar la pendiente de las curvas de nivel de la función $\frac{\widehat{I}_R(\omega)}{\widehat{I}_S(\omega)} = e^{-i\omega^T x_0}$, lo cual puede realizarse mediante la aplicación de técnicas de regresión [SOC99].

Un método de registro bastante robusto, basado también en la transformada de Fourier, y llamado *detección de traslación por restauración* [VSOB99], se deduce al multiplicar ambos lados de la ecuación (4.4) por $\widehat{I}_S^*(\omega)$ y dividir posteriormente por $|\widehat{I}_S(\omega)|^2$:

$$\frac{\widehat{I}_R(\omega)\widehat{I}_S^*(\omega)}{|\widehat{I}_S(\omega)|^2 + \mu} = e^{-\omega^T x_0} \ , \tag{4.9}$$

el término μ es una constante que se agrega al denominador para considerar los efectos del ruido. De manera similar a los métodos de registro basados en la correlación de fase, el vector de traslación se encuentra maximizando con respecto a la variable espacial la transformada inversa de Fourier del término izquierdo de la ecuación (4.9).

El registro de imágenes mediante la transformada de Fourier puede extenderse para recuperar transformaciones compuestas por rotaciones y escalamientos, además de traslaciones, considerando la representación en coordenadas polares de la transformación de Fourier, la cual es conocida como el descriptor Fourier-Mellin [CDD94, KSA05, GXLP05, PQC08]. En estas circunstancias, el efecto de la rotación de la imagen I_S se refleja en un desplazamiento del argumento angular del descriptor Fourier-Mellin; el ángulo de rotación puede obtenerse aplicando el método de corrrelación de fase a la magnitud del descriptor Fourier-Mellin de las imágenes originales.

Los métodos de registro de imágenes basados en la transformación de Fourier permiten recuperar traslaciones, rotaciones y escalamientos de gran magnitud, con un bajo costo computacional, además de que algunos de ellos son robustos ante cambios de iluminación entre cuadros y a la presencia de ruido.

Sin embargo, su principal desventaja radica en la dificultad que existe para aplicarse en problemas de registro bajo transformaciones más generales.

4.3. Registro de Imágenes mediante Medidas de Información.

La aplicación de la información mutua en el registro de imágenes fue introducida simultáneamente en los trabajos de Viola *et al* [VWI95] y Collignon *et al* [CMD+95a], a mediados de la década pasada. En ambos trabajos el registro se lleva a cabo encontrando la transformación T que maximiza el valor de la información mutua entre la imagen transformada y la imagen de referencia; esto es, se busca T^* tal que:

$$T^* = \arg\max_T IM(T) = H(I_T) + H(I_R) - H(I_T, I_R) \ . \tag{4.10}$$

La idea intuitiva detrás de esta metodología consiste en que la entropía de la distribución $p(I_T, I_S)$ es mínima cuando ambas imágenes están alineadas, dado que al coincidir espacialmente las estructuras correctas se generan grupos de alta densidad sobre sus tonos de gris. Al desalinear las imágenes, los puntos pertenecientes a una estructura en particular, se traslapan sobre diferentes regiones en la imagen complementaria, lo que da lugar a la dispersión de los grupos de alta densidad en la distribución conjunta, aumentando su entropía. Al mismo tiempo que se minimiza la entropía conjunta se busca que la transformación mantenga información (estructura) en el traslape de ambas imágenes maximizando $H(I_T) + H(I_R)$; lo anterior debido a que una manera trivial de minimizar la entropía conjunta es hacer que las imágenes coincidan en regiones sin estructura (imágenes que se traslapan en un sólo

punto por ejemplo).

La técnica de registro de imágenes por maximización de información mutua ha sido ampliamente adoptada desde su introducción y su aplicación se ha extendido a la solución de otros problemas como la segmentación de pares estereoscópicos [JK03], el seguimiento (tracking) de rasgos [DB08], la restauración [CWFT05] y la segmentacón de imágenes [SD06]; convirtiéndose en una elección muy adecuada ante la necesidad de integrar información proveniente de diversas fuentes.

Resulta interesante realizar un análisis de sensibilidad para la entropía, considerando que ésta es una función de las entradas del vector de probabilidades. Expandiendo (2.3) mediante una serie de Taylor se obtiene la siguiente expresión:

$$\Delta H(\mathbf{p}) = -\sum_i \left[1 + \log p_i\right] \Delta p_i \tag{4.11}$$

siendo $p_i = p(X = x_i)$. Si se parte de un valor fijo de entropía y se realiza una actualización del vector de probabilidades sumando Δp_i a cada p_i, la entropía se modificará de acuerdo a (4.11), y debido a que el coeficiente $[1 + \log p_i]$ tiene una gran magnitud para las entradas del vector de probabilidad más pequeñas, éstas tienen un efecto importante sobre el incremento de entropía. Al registrar un par de imágenes mediante (4.10) utilizando algún método de optimización iterativo, las actualizaciones en información mutua están determinadas directamente por los incrementos en las entropías marginales y conjunta, y éstos reflejan fuertemente los cambios en los rasgos

menos importantes de las imágenes (al estar relacionados generalmente con valores de probabilidad muy pequeños). Como consecuencia, el proceso de optimización puede enfrentarse a la presencia de múltiples óptimos locales, los cuales se generan al alinear temporalmente rasgos poco importantes en las imágenes. En general, es posible relizar el registro a través de estrategias multiescala (coarse to fine) para reducir esta sensibilidad, sin embargo esto aún representa una fuerte desventaja en los casos en que es necesario aplicar transformaciones de gran magnitud para alinear un par de imágenes. Lo anterior contrasta con la kérnel-predictibilidad, ya que como se demostró anteriormente (ver sección 3.3), los incrementos de esta medida se encuentran determinados fuertemente por las entradas de mayor magnitud del vector de probabilidades.

En la práctica, además, es importante tomar en cuenta que los resultados obtenidos con (4.10) son sensiblemente afectados por diferentes estrategias de implementación [ZC02]. Entre los factores más importantes se encuentran, la estimación de las distribuciones de probabilidad, el manejo del traslape entre las imágenes, la optimización y la interpolación utilizada para estimar intensidades en puntos de la imagen transformada. Estos puntos se describirán en las siguientes subsecciones.

4.3.1. Estimación de las Distribuciones de Probabilidad.

En la aplicación de (4.10) es necesario estimar la distribución de probabilidad conjunta $\mathbf{p}(I_T, I_R)$, y las distribuciones marginales $\mathbf{p}(I_T)$ y $\mathbf{p}(I_R)$, para lo cual, tradicionalmente han sido empleadas dos estrategias. La primera de

ellas consiste en construir el histograma conjunto sobre la región de traslape de las dos imágenes y después normalizar cada entrada, con lo que se obtiene:

$$\mathbf{p}\left(I_T = a, I_R = b\right) = \frac{1}{|Tr|} \sum_{\mathbf{x} \in Tr} \delta\left[I_T(\mathbf{x}) - a\right] \delta\left[I_R(\mathbf{x}) - b\right] \tag{4.12}$$

en donde a es cualquier valor del conjunto de tonos de gris de la imagen fuente y b de la imagen de referencia. La suma se extiende sobre los puntos ubicados en la región de traslape de las dos imágenes, Tr y $\delta(\cdot)$ es la función delta de Kronecker. Las distribuciones marginales se obtienen sumando sobre los renglones o columnas de $\mathbf{p}(I_T, I_R)$. Posteriormente, los valores de entropía son evaluados utilizando (2.3).

A pesar de que la construcción de los histogramas normalizados puede realizarse con un bajo costo computacional, para obtener una buena aproximación a la entropía es necesario disponer de un número suficiente de muestras; condición que puede dificultar el trabajo con imágenes bidimensionales, sobre todo al realizar el registro mediante estrategias multiescala, en donde generalmente se registran versiones reducidas de las imágenes originales obtenidas por submuestreo. En esas condiciones es necesario también cuantizar el número de tonos de gris (entradas del histograma) de manera que se refleje la resolución de las imágenes en cada escala: en las mayores escalas se usan pocos bines, mientras que en las escalas menores se utiliza un número mayor [ZC02]. Otra desventaja radica en que la expresión (4.12) no es diferenciable, lo cual limita las opciones de optimización para maximziar la información mutua.

La segunda estrategia realiza una estimación no-paramétrica de las densidades mediante ventanas de Parzen; estas ventanas se centran sobre un conjunto de valores de intensidad muestreados en el traslape de las imágenes. Un nuevo conjunto de muestras permite aproximar la entropía evaluando el promedio de la información de la distribución:

$$H(\mathbf{I}) = -\frac{1}{|A|}\sum_{i\in A}\log\left[\frac{1}{|B|}\sum_{j\in B}G_\sigma\left(\|\mathbf{I}(x_i)-\mathbf{I}(x_j)\|\right)\right] \tag{4.13}$$

en donde $G_\sigma()$ es una ventana de Parzen, generalmente gaussiana con parámetro de amplitud σ. A y B son los conjuntos de muestras con valores de intensidad del vector $\mathbf{I} = (I_R, I_T)$. Las entropías marginales se evalúan con la misma expresión, sustituyendo $\mathbf{I}$ por I_T e I_R. Las ventanas de Parzen están sujetas a las siguientes restricciones:

$$\begin{aligned}\int_{-\infty}^{\infty} G(x)dx &= 1\\ G(x) &\geq 0\ ,\ \forall x.\end{aligned}$$

Esta estrategia permite trabajar con un número menor de muestras que las necesarias para construir el histograma normalizado. Aunado a esto, la expresión (4.13) es derivable si la ventana de Parzen también lo es. No obstante estas ventajas, el costo computacional de (4.13) es cuadrático respecto al número de muestras.

Otras alternativas han sido propuestas para la estimación de las distribuciones de probabilidad. En el caso de imágenes bidimensionales, Rajwade *et al* [RBR06, RBR08], calculan el valor $p(I = \alpha)$ integrando el inverso

de la magnitud del gradiente de la imagen sobre las curvas de nivel α, ellos derivan este resultado al considerar que $p(I = \alpha) = \frac{d}{d\alpha}[F_I(\alpha)]$ y que $F_I(\alpha) = \frac{\int\int_{I(x,y)<\alpha} dxdy}{\int\int dxdy}$; la densidad conjunta es calculada de manera similar. Dowson *et al* [DKB08], construyen una versión continua de las imágenes mediante interpolación bilineal y calculan la distribución de probabilidad en cada sección de interpolación (localizada entre 4 puntos adyacentes) mediante el uso de fórmulas de transformación estándares; la distribución global se obtiene considerando las aportaciones de todas las secciones de interpolación.

4.3.2. Interpolación.

Construir la imagen transformada I_T implica evaluar la intensidad de I_S en cualquier valor continuo, mientras que la imagen se encuentra definida en un conjunto discreto de puntos, por lo cual es necesario utilizar algún tipo de interpolación. Diferentes alternativas han sido propuestas, siendo las más simples la interpolación por vecinos cercanos y la lineal. Estos métodos, sin embargo, pueden producir cambios irregulares en los histogramas de intensidad, aún al variar suavemente la imagen transformada, como consecuencia de posibles alineamientos temporales de grandes volúmenes de puntos en la malla de la imagen, lo cual complica el registro [MVS99, CMD+95b]. Tratando de evitar este problema, Collignon *et al* [CMD+95b], proponen el uso de *interpolación de volumen parcial*, en este método los valores de intensidad de la imagen transformada son calculados mediante interpolación lineal, mientras que la actualización de los histogramas se realiza considerando las entradas de todos los vecinos del valor de intensidad interpolado, en lugar de actualizar únicamente la entrada correspondiente, con lo que se obtienen histogramas

normalizados más suaves; cabe señalar que Collignon utiliza histogramas con un número fijo de entradas durante todo el proceso de registro, mientras que Zhu [ZC02], no encuentra desventajas en utilizar interpolación lineal normal siempre y cuando el registro se realice bajo alguna estrategia multiescala, con histogramas de pocas entradas en escalas bajas y de muchas entradas en las escalas altas. Buscando evitar los artefactos introducidos por la interpolación lineal, Likar [LP01] en lugar de interpolar las intensidades de las imagen transformada sobre el punto $T(x)$ de I_S, lo hace en la posición $T(x) + \delta x$, siendo δx un valor aleatorio generado (para cada x) de una distribución uniforme de pequeña amplitud. Thevenaz *et al* [TU00], reportan una mejora en el desempeño del registro con el uso de interpolación de alto orden, particularmente mediante splines cúbicos, sin embargo la utilización de este tipo de interpolación incrementa el costo computacional del registro.

Se debe notar que la estrategia de interpolación puede ser un factor crítico al estimar las densidades de probabilidad mediante histogramas normalizados, mientras que, dentro de los límites de la revisión bibliográfica realizada, no se encontró ningún reporte acerca de estos problemas al utilizar ventanas de Parzen.

4.3.3. Otras Medidas de Información.

La entropía de Shannon (2.3) no es la única medida de información que ha sido aplicada en el registro de imágenes. Rodríguez y Loew [RCL98], proponen el uso de la entropía de Jumarie, extendiendo el problema de registro al considerar las coordenadas espaciales y la intensidad de cada punto de

la imagen y su vecindad; esta idea sin embargo incrementa fuertemente la dimensionalidad del problema dificultando su implementación. En [BFB04], Bardera *et al* proponen la utilización de la entropía de Tsallis. Mientras que Zhang y Rangarajan [ZR04], utilizan la suma de entropías condicionales, $H(I_T|I_R) + H(I_R|I_T)$, como una nueva medida de similitud entre imágenes, esta nueva medida cumple con más propiedades de métrica que la información mutua, presentando algunas ventajas sobre ésta al aplicarse al registro simultáneo de múltiples imágenes.

Studholme *et al*, [SHH99], observaron que al registrar imágenes médicas con un campo de visión muy grande (imágenes de fondo muy amplio comparado con la región de interés), la información mutua entre las imágenes puede llegar a incrementarse, en lugar de disminuir, al alejar la transformación de su valor óptimo. Para superar este fenómeno, Studholme propuso la utilización de la siguiente medida, la cual es conocida como *Información Mutua Normalizada*:

$$IMN(T) = \frac{H(I_T) + H(I_R)}{H(I_T, I_R)} \ . \tag{4.14}$$

Esta medida tiene además la propiedad de ser invariante ante la magnitud del traslape entre las imágenes.

La información mutua como medida de similitud permite alinear imágenes independientemente de la relación entre sus tonos de gris, por lo cual es un método con aplicaciones muy generales. En el procesamiento de imágenes médicas sin embargo, las modalidades de las imágenes a registrar no son

arbitrarias; la necesidad de integrar información se presenta en unos cuantos pares de modalidades, por ejemplo, imágenes de resonancia magnética e imágenes del tipo PET o SPECT. Al registrar frecuentemente este tipo de imágenes se puede obtener conocimiento de la distribución conjunta de las imágenes alineadas y utilizarla como referencia para guiar el registro. Siguiendo esta idea, Chan [CCY+03], propone el registro de imágenes mediante la minimización de la distancia de Kullback-Leibler entre la distribución conjunta de las imágenes a registrar y la de imágenes alineadas de las mismas modalidades obtenida a priori. La distancia de Kullback-Leibler es la siguiente:

$$KLD(P_1|P_2) = \sum_{i,j} P_1(X_i, X_j) \log \left[\frac{P_1(X_i, X_j)}{P_2(X_i, X_j)} \right] , \tag{4.15}$$

siendo P_1 la distribución de las imágenes a registrar y P_2 la distribución aprendida.

En lugar de la distancia de Kullback-Liebler, Sun y Guo [SG07] utilizan la medida de divergencia de Tsallis:

$$TDM(P_1|P_2) = \frac{1}{\alpha - 1} \left[1 - \sum_{i,j} \left(\frac{P_1^{\alpha}(X_i, X_j)}{P_2^{\alpha-1}(X_i, X_j)} \right) \right] . \tag{4.16}$$

4.3.4. Optimización

El éxito del registro a través de la maximización de la información mutua depende fuertemente del método de optimización elegido. La estimación de las distribuciones de probabilidad es el factor que determina principalmente la elección. Como se mencionó anteriormente, la utilización de his-

togramas normalizados no permite obtener una expresión diferenciable, por lo que Collignon [CMD+95b], realiza la maximización mediante el método de Powell [PTVP99]. Para maximizar una función $f(\mathbf{x}) = f(x_1, x_2, \ldots, x_n)$, el método de Powell hace uso de un conjunto compuesto por n vectores de dirección, $D = \{d_1, d_2, \ldots, d_n\}$, y partiendo de un valor inicial, x_0, genera una secuencia de puntos, $x_0 = p_0, p_1, \ldots, p_n = x_1$, en donde el punto p_i maximiza la función original en la dirección d_i, esto es, $p_i = p_{i-1} + \gamma_i d_i$, siendo $\gamma_i = \arg \max_\gamma f(p_{i-1} + \gamma d_i)$. La optimización de la función original en una dirección cualquiera (un problema de optimización unidimensional) puede realizarse a través del método de la razón aurea (golden ratio) [PTVP99], el cual trata de acotar la presencia del óptimo recortando iterativamente un intervalo inicial. Una vez exploradas todas las direcciones del conjunto D, el punto actual se mueve a x_1 y los vectores de dirección se actualizan haciendo $d_i = d_{i+1}$ para $i < n$ y $d_n = p_n - p_0$. El proceso de exploración de todas las direcciones del conjunto D se repite iterativamente ingresando en cada paso el vector $p_n - p_0$ y eliminando el vector d_1. A pesar de que el método Powell no requiere el cálculo de derivadas es muy sensible a la presencia de óptimos locales. Bajo este mismo enfoque, Zhu [ZC02] reporta mejores condiciones de convergencia hacia la transformación óptima optimizando con el método Simplex down-hill [PTVP99]. Este método se basa en la construcción de un simplejo en el espacio n dimensional, el cual se actualiza iterativamente a través de las operaciones de contracción, contracción múltiple, expansión y reflexión aplicadas a sus vértices. Considerando los métodos que utilizan ventanas de Parzen, Viola [VWI95] optimiza con ascenso de gradiente estocástico (renovando los conjuntos de muestreo en cada

cálculo del gradiente), lo cual presenta cierta robustez ante la presencia de óptimos locales. Thévenaz y Unser [TU00] presentan una estrategia de optimización basada en el método de Levenberg-Marquardt. Algunos autores proponen la utilización de métodos basados en algoritmos evolutivos, Butz y Thiran realizan el registro maximizando la información mutua mediante algoritmos genéticos distribuídos [BT01]. Los algoritmos genéticos también son utilizados en [ZZSZ05, YMLL07]. Gómez *et al* [Gar02, GVA$^+$02] utilizan estrategias evolutivas $(\mu + \lambda)$; mientras que el método de evolución diferencial, implementado en un algoritmo paralelo, es explorado por De Falco *et al* [FMST07].

4.3.5. Manejo del Traslape

La estimación de la información mutua, debe realizarse muestreando sobre la región de traslape de ambas imágenes. Esta región, está determinada por la proyección de la imagen transformada I_T sobre la imagen I_R, para un valor de T dado. Al actualizar la transformación, la parte de I_R ubicada en la región de traslape puede cambiar, por lo que, en general, la entropía de I_R no permanecerá constante durante el proceso de registro. No obstante que existe una dependencia entre $H(I_R)$ y la transformación, ésta no puede describirse de forma explícita, siendo un factor que dificulta el cálculo del gradiente de la información mutua. Obviamente este problema afecta únicamente a los métodos de registro que utilicen derivadas de la información mutua. Para evitar esta dificultad, Viola *et al* [VWI95] asignan arbitrariamente valores de intensidad cero a los puntos $T(x)$, ubicados fuera de los límites originales de la imagen I_S, lo que equivale a extender infinitamente esta imagen con un fondo

de intensidad cero; bajo estas condiciones la región de traslape entre las dos imágenes es siempre la totalidad de I_R y por consiguiente $H(I_R)$ permanece constante. Esta solución, aunque parece ser apropiada para algunas imágenes (sobre todo algunas imágenes médicas que por su naturaleza ya presentan un fondo con intensidad cero), puede aumentar el campo de visión de la imagen original, lo cual, como se mencionó anteriormente, dificulta el proceso de registro. Una mejor opción consiste en aproximar las derivadas mediante diferencias finitas, las cuales son exploradas en este trabajo.

4.4. Otras Medidas de Similitud entre Imágenes.

Además de la información mutua, otras medidas de similitud entre imágenes han sido aplicadas al problema de registro. Particularmente, la necesidad de alinear imágenes médicas multimodales ha motivado la exploración de diferentes medidas de similitud, algunas de las cuales presentan propiedades que resultan ventajosas en aplicaciones específicas. No obstante, cabe hacer notar que el registro de imágenes basado información mutua, continúa siendo la técnica más ampliamente utilizada, dada la generalidad de sus aplicaciones. Diferentes comparaciones [SLP04] muestran la ventaja de utilizar la información mutua (y su versión normalizada) como medida de similitud entre imágenes, con respecto a algunas de las medidas que serán descritas en esta sección.

4.4.1. Correlación Cruzada.

Esta medida se utiliza principalmente para ubicar la posición de patrones dentro de una imagen dada [Jan02]. Cualquier patrón, S, queda determinado

por sus valores de intensidad dentro de una región de definición Ω, esto es, $S : \Omega \rightarrow R$. El patrón es equivalente a un vector real cuya dimensión es igual a la cardinalidad de Ω (considerando que tanto el patrón como la imagen se definen en conjuntos discretos). La ubicación de S en la imagen dada I, se realiza encontrando la posición que maximiza la correlación cruzada entre el patrón y la imagen:

$$x^* = \arg \max_x \frac{\sum_{x'} S[x']I[x' - x]}{\sqrt{\left(\sum_{x'} S^2[x']\right)\left(\sum_{x'} I^2[x' - x]\right)}}$$

La correlación cruzada, evaluada sobre un punto x, es equivalente al producto punto normalizado entre el patrón y el vector que resulta al considerar la intensidad de los puntos de I ubicados en el traslape con Ω, después de trasladar la imagen I para hacer coincidir el punto x sobre el origen de la región Ω. Dado que el producto punto está normalizado, la correlación cruzada es máxima, e igual a uno, cuando los dos vectores son paralelos, por lo que esta medida permite identificar patrones dentro de imágenes aún cuando exista un factor multiplicativo entre las intensidades del patrón y la imagen.

Es bien conocido el hecho de que la correlación cruzada presenta una multiplicidad de óptimos locales, además de que, por definición, esta medida puede aplicarse únicamente en el registro de imágenes bajo traslaciones constantes, por lo que su utilidad es limitada.

4.4.2. Coeficiente de Correlación.

El coeficiente de correlación, definido en la ecuación (2.1), mide la dependencia lineal entre un par de variables aleatorias. Si existe una relación

lineal entre ambas variables entonces el valor absoluto del coeficiente de correlación es igual a uno, mientras que si ambas variables son independientes el coeficiente es cero.

La utilización de esta medida en el registro de imágenes ha sido explorada en los siguientes trabajos [JMH+90, BDC+93, BGL+93, EP08]; sin embargo, su aplicación es limitada, ya que en general las dependencias entre las intensidades de las imágenes pueden llegar a ser más generales que una relación lineal.

4.4.3. Razón de Correlación

Es posible extender el concepto de coeficiente de correlación para considerar relaciones más generales. Si suponemos que $Y = \Phi(X)$ (siendo Φ desconocida), entonces la calidad de cualquier función Ψ para aproximar esta dependencia puede cuantificarse por medio de algún funcional de costo. Más aún, la mejor estimación para la función Φ, se encuentra minimizando el valor esperado de este costo:

$$\begin{aligned} \widehat{\Phi} &= \arg\min_{\Psi} E\left\{L(Y - \Psi(X))\right\} \\ &= \arg\min_{\Psi} \int_Y \int_X L\left[y - \Psi(x)\right] p(x,y) dx dx y \ . \end{aligned} \tag{4.17}$$

Si L es un funcional cuadrático ($L[Y - \Psi(X)] = [Y - \Psi(X)]^2$), entonces la energía (4.17) se minimiza para $\widehat{\Phi}(x) = E(Y|X = x)$ [Bis06].

Una vez definida la función $\widehat{\Phi}$, la varianza de la variable Y puede descomponerse de la siguiente manera:

$$Var(Y) = Var(E(Y|X)) + E_X Var(Y|X = x) \tag{4.18}$$

Mientras que el primer término del lado derecho de (4.18) mide la cantidad de la varianza de la variable Y que es explicada por el modelo $Y = \widehat{\Phi}(X)$, el segundo cuantifica la varianza de la parte de Y que es funcionalmente independiente de X. En base a (4.18), la razón de correlación se define como la proporción de la varianza de Y que es explicada por el modelo [RMPA98]:

$$\begin{aligned} CR(X,Y) &= \frac{Var(E(Y|X))}{Var(Y)} \\ &= 1 - \frac{E_X Var(Y|X = x)}{Var(Y)} \, . \end{aligned}$$

El registro de imágenes puede realizarse encontrando la transformación que maximice la razón de correlación entre la imagen transformada y la imagen de referencia. Aunque en la práctica suele minimizarse la expresión:

$$\begin{aligned} \frac{E_X Var(Y|X = x)}{Var(Y)} &= 1 - CR(X,Y) \\ &= \frac{\int_X Var(Y|X = x) p(X = x) dx}{Var(Y)} \, . \end{aligned} \tag{4.19}$$

Debe notarse que la razón de correlación no es una medida simétrica; esta característica es importante cuando la relación funcional entre las intensidades de las imágenes a registrar no es invertible, en esos casos la imagen a transformar debe seleccionarse asegurando que la distribución $p(I_T|I_R = i)$ sea monomodal para todo valor de i. Lo anterior representa una desventaja

importante para la aplicación general de la razón de correlación en el registro multimodal de imágenes.

5. REGISTRO DE IMÁGENES MEDIANTE KERNEL-PREDICTIBILIDAD

La aplicación de KP al problema de registro de imágenes puede realizarse considerando la distribución conjunta de las intensidades de las imágenes I_R e I_T. Definiendo $\mathbf{I}_J(T) \equiv < I_R, I_T >$, es posible reescribir esta distribución como $p(\mathbf{I}_J(T))$. La idea intuitiva se basa en que cuando $T = T^*$, en donde T^* es la transformación que alinea correctamente las imágenes, $p(\mathbf{I}_J(T^*))$ debe tener una menor dispersión que $p(\mathbf{I}_J(T))$ para $T \neq T^*$, y por lo tanto $KP[p(\mathbf{I}_J(T^*))] > KP[p(\mathbf{I}_J(T))]$ para $T \neq T^*$. Por ejemplo, si suponemos que existe una función de transferencia de tonos Φ, entre I_R e I_{T^*}, $p(\mathbf{I}_J(T^*))$ deberá ordenarse a lo largo de una estructura determinada por Φ: en este caso, la distribución condicional $p(I_{T^*}|I_R = i) = \delta(I_{T^*} - \Phi(i))$, y cualquier otra transformación debe redistribuir la densidad condicional en diferentes valores de tonos de gris.

Sin embargo, no es suficiente considerar únicamente el valor de KP evaluado sobre la distribución conjunta de I_R e I_T, dado que por ejemplo, este valor puede maximizarse bajo transformaciones que asignen todos los puntos en la imagen I_S a un único punto en I_R. Para evitar esta situación, el espacio de soluciones debe restringirse; siendo una opción la normalización del valor

de KP conjunto, la cual se analiza en la siguiente sección.

5.1. Medidas de Similitud entre Imágenes Basadas en Kérnel-Predictibilidad

El valor de KP conjunto puede normalizarse dividiéndolo entre la suma de las KP marginales de una manera similar a como se realiza con la información mutua normalizada [SHH99]. Lo anterior deriva en la siguiente medida de similitud entre imágenes:

$$SKP_1(I_T, I_R) = \frac{KP[p(\mathbf{I}_J)]}{KP[p(I_T)] + KP[p(I_R)]} \ . \tag{5.1}$$

Esta medida de similitud puede describirse como una comparación entre la predictibilidad de la distribución conjunta y la predictibilidad de las distribuciones marginales de las imágenes I_T e I_R.

En el caso discreto es posible derivar una cota superior para la SKP_1. Supongamos que para evaluar la KP de un par de imágenes I_T e I_R, se utiliza un kérnel K, con la siguiente propiedad: $K(i,i) = 1 \geq K(i,j)$, para $i \neq j$. Baándose en este kérnel puede construirse un kérnel separable, K_J, para medir la KP de la distribución conjunta, $p(\mathbf{I}_J)$, como:

$$K_J((i_1, j_1), (i_2, j_2)) = K(i_1, i_2)K(j_1, j_2) \tag{5.2}$$

se tiene entonces:

$$\begin{aligned} KP[p(\mathbf{I}_J)] &= \sum_{i_1}\sum_{i_2}\sum_{j_1}\sum_{j_2} p_{\mathbf{I}_J}(i_1,j_1)p_{\mathbf{I}_J}(i_2,j_2)K_J((i_1,j_1),(i_2,j_2)) \\ &= \sum_{i_1}\sum_{i_2} p_{I_T}(i_1)p_{I_T}(i_2)K(i_1,i_2)\sum_{j_1}\sum_{j_2} p_{I_R}(j_1|i_1)p_{I_R}(j_2|i_2)K(j_1,j_2) \\ &\leq \sum_{i_1}\sum_{i_2} p_{I_T}(i_1)p_{I_T}(i_2)K(i_1,i_2) = KP[p(I_T)] \end{aligned} \tag{5.3}$$

De manera similar puede verse que $KP[p(\mathbf{I}_J)] \leq KP(p(I_R))$, por lo tanto $SKP_1(I_T, I_R) \leq \frac{1}{2}$.

Si se considera ahora una imagen de referencia I_R y una imagen transformada I_T, y se asume que cuando se aplica la transformación que alinea correctamente ambas imágenes (T^*), las intensidades i_R, i_{T^*} se relacionan por medio de una función de transferencia de tonos invertible Φ, entonces $p(i_{T^*}|i_R) = \delta(i_{T^*} - \Phi(i_R))$. Si se asume además que $K(i,j) = \delta(i-j)$, entonces, de (5.3) puede verse que $KP(p(I_R, I_{T^*})) = KP(p(I_R)) = KP(p(I_{T^*}))$, y por lo tanto $SKP_1(I_R, I_{T^*}) = \frac{1}{2}$. Lo anterior significa que SKP_1 alcanza su máximo global cuando $T = T^*$.

En el caso del kérnel gaussiano, por lo menos un máximo local de SKP_1 se obtiene bajo la transformación óptima; esto se deriva al notar que para una transformación T diferente, pero cercana a T^*, $KP[p(I_T)]+KP[p(I_R)] \approx KP[p(I_{T^*})]+KP[p(I_R)]$ y $KP[p(\mathbf{I}_J(T^*))] > KP[p(\mathbf{I}_J(T))]$, dado que $p(\mathbf{I}_J(T))$ tiene una concentración menor que $p(\mathbf{I}_J(T^*))$ (considerar la ecuación (3.7) y la discusión anterior). En la práctica se obtiene también un máximo local

bajo la transformación T^* empleando kérneles suaves, para los cuales KP tiene un comportamiento similar al del caso gaussiano (ver sección 3.1).

La estimación de la KP conjunta a través de un conjunto de muestreo y utilizando kérneles separables, ecuación (5.2) es equivalente al producto punto de dos vectores, como puede verse en la siguiente ecuación:

$$\begin{aligned} \widehat{KP}\left[p(\mathbf{I}_J)\right] &= \sum_i \sum_j K_2(\mathbf{I}_J^i, \mathbf{I}_J^j) \\ &= \sum_i \sum_j K(I_R^i, I_R^j) K(I_T^i, I_T^j) \\ &= \langle \mathbf{K}(I_R), \mathbf{K}(I_T) \rangle \end{aligned} \tag{5.4}$$

en donde los índices permiten recorrer diferentes pares de muestras dependiendo del estimador utilizado e $\mathbf{I}_J^i = (I_T^i, I_R^i) = (I_T(X_i), I_R(X_i))$

Una nueva medida de similitud se deriva al considerar la normalización de este producto punto:

$$\begin{aligned} SKP_2(I_T, I_R) &= \frac{\sum_i \sum_j K(I_R^i, I_R^j) K(I_T^i, I_T^j)}{\sqrt{\sum_i \sum_j K^2(I_R^i, I_R^j) \sum_i \sum_j K^2(I_T^i, I_T^j)}} \\ &= \frac{\langle \mathbf{K}(I_R), \mathbf{K}(I_T) \rangle}{\|\mathbf{K}(I_R)\| \|\mathbf{K}(I_T)\|} \, . \end{aligned} \tag{5.5}$$

Tomando en cuenta la desigualdad de Cauchy-Schwartz, $|SKP_2(I_T, I_R)| \leq 1$, independientemente de la selección de K. Cuando las imágenes I_T e I_R se encuentran perfectamente alineadas ($T = T^*$) y existe entre ellas una función

de transferencia de tonos igual a la identidad, se tiene que $I_T^i = I_R^i$ e $I_T^j = I_R^j$, y por lo tanto esta nueva medida de similitud adquiere su máximo global: $SKP_2(I_{T^*}, I_R) = 1$.

En la práctica los resultados obtenidos mediante la utilización de SKP_1 y SKP_2 son muy similares (tal como se muestra en la figura 5.5). Sin embargo la evaluación de SKP_2 conlleva una carga computacional más alta que la necesaria para calcular SKP_1, dado que se requiere elevar al cuadrado cada uno de los elementos de los dos vectores; aunque esta diferencia desaparece si se tabulan previamente las evaluaciones de los kérneles. Esta estrategia además acelera la ejecución de los algoritmos de registro significativamente, sobre todo al utilizar kérneles muy complejos de evaluar. En este trabajo se exploró la utilización de SKP_1, por lo que, en adelante, las siglas SKP se utilizarán para referirse específicamente a SKP_1 (a menos que por el contexto sea necesario diferenciar entre ambas medidas de similitud).

El registro de las imágenes I_S e I_R se realiza buscando la transformación T que maximice el valor de SKP entre las imágenes I_T e I_R. La transformación puede clasificarse como *paramétrica* o *no-parametrica*; una estrategia de registro diferente debe seguirse en cada caso, como se detalla en las secciones 5.3 y 5.4. Asumiendo que, por el contexto, es claro sobre qué imágenes se evalúa la medida de similitud, se escribirá $SKP(T)$ en lugar de la expresión $SKP(I_T, I_R)$.

5.2. Relación de SKP con Otras Medidas de Similitud.

En [BFB04], Bardera *et al* muestran que al utilizar la entropía de Tsallis bajo un enfoque similar al de la información mutua normalizada (4.14) se genera una medida de similitud entre imágenes con una gran robustez. Específicamente proponen el empleo de la entropía de Gini, obtenida de la entropía de Tsallis (2.6) seleccionando el parámetro libre igual a dos, dada su sencillez computacional y su alta convergencia hacia la transformación óptima, demostrada en el registro de algunas modalidades de imágenes médicas. Como se mostró en la sección 3.1, al utilizar un kérnel igual a la delta de Kroneker el valor de KP de un vector de probabilidades es equivalente al negativo de la entropía de Gini del mismo vector (más una unidad), por lo que, para ese kérnel específico, SKP presenta una gran similitud con la medida propuesta por Bardera *et al*.

Por otro lado, haciendo $K_R(I_R^i, I_R^j) = \delta(I_R^i - I_R^j)$, $K_T(I_T^i, I_T^j) = -(I_T^i - I_T^j)^2$ y $K_J = K_R(I_R^i, I_R^j)K_T(I_T^i, I_T^j)$, el valor de SKP, entre un par de imágenes I_T e I_R, se reduce a:

$$SKP(T) = -\frac{\int_X Var(I_T|I_R = x)p^2(I_R = x)dx}{Var(I_T) + \mu}$$

en donde $\mu = [1 - Gini(p(I_R))]$, es una constante. Debe notarse la similitud entre esta medida y la razón de correlación, definida en la ecuación (4.19). Sin embargo, deben recordarse también las limitaciones de la varianza para medir la dispersión de distribuciones multimodales (sección 2.2), lo que explica la dificultad para utilizar la razón de correlación en problemas de registro multimodal de imágenes en general.

5.3. Registro Paramétrico

Supongamos que la trasformación T se determina por un vector de m parámetros reales $\mathbf{a} = (a_1, a_2, \ldots, a_m)$, y que m es considerablemente menor que el número total de puntos sobre las imágenes a registrar; en este caso, podemos escribir $T(x; \mathbf{a})$ en lugar de $T(x)$, por ejemplo al registrar imágenes bajo transformaciones afines o proyectivas. Entonces una aproximación a (5.1) usando el estimador (3.11) puede escribirse de la siguiente manera:

$$\widehat{SKP}[T(\mathbf{a})] = \frac{\widehat{KP}_J[T(\mathbf{a})]}{\widehat{KP}_T[T(\mathbf{a})] + \widehat{KP}_R} \tag{5.6}$$

con

$$\begin{aligned}
\widehat{KP}_J[T(\mathbf{a})] &= \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} K_J(\mathbf{I}_J^i, \mathbf{I}_J^j) \\
\widehat{KP}_T[T(\mathbf{a})] &= \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} K_M(I_T^i, I_T^j) \\
\widehat{KP}_R &= \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} K_M(I_R^i, I_R^j) \ ,
\end{aligned}$$

y en donde K_J es el kérnel utilizado para medir la predictibilidad de la distribución conjunta de I_T e I_R; K_M para las distribuciones marginales de I_T e I_R. Nótese que el coeficiente constante en los estimadores puede ignorarse debido a la normalización.

En caso de emplear kérneles gaussianos (e ignorando la constante de normalización):

$$K_J(\mathbf{I}_J^i, \mathbf{I}_J^j) = G_{\sigma_J}(\mathbf{I}_J^i, \mathbf{I}_J^j) = \exp\left\{-\frac{\|\mathbf{I}_J^i - \mathbf{I}_J^j\|^2}{2\sigma_J^2}\right\} \quad (5.7)$$

$$K_M(I^i, I^j) = G_{\sigma_M}(I^i, I^j) = \exp\left\{-\frac{(I^i - I^j)^2}{2\sigma_M^2}\right\} . \quad (5.8)$$

La maximización se realiza utilizando ascenso de gradiente estocástico, iniciando con una transformación definida por el vector $\mathbf{a}^0$, y actualizándola mediante la relación:

$$\mathbf{a}^{t+1} = \mathbf{a}^t + \lambda \nabla_{\mathbf{a}} \widehat{SKP}\left[T(\mathbf{a}^t)\right]$$

con:

$$\nabla_{\mathbf{a}} \widehat{SKP}\left[T(\mathbf{a}^t)\right] = \frac{1}{\widehat{KP}_T\left[T(\mathbf{a}^t)\right] + \widehat{KP}_R} \nabla_{\mathbf{a}} \widehat{KP}_J\left[T(\mathbf{a}^t)\right] - \frac{\widehat{KP}_J\left[T(\mathbf{a}^t)\right]}{(\widehat{KP}_T\left[T(\mathbf{a}^t)\right] + \widehat{KP}_R)^2} \nabla_{\mathbf{a}} \widehat{KP}_T\left[T(\mathbf{a}^t)\right] \quad (5.9)$$

y en particular, al emplear los kérneles gaussianos (5.7) (5.8), estos gradientes pueden escribirse:

$$\nabla_{\mathbf{a}} \widehat{KP}_J\left[T(\mathbf{a}^t)\right] = -\frac{1}{\sigma_J^2} \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} G_{\sigma_J}(\mathbf{I}_J^i, \mathbf{I}_J^j)(I_T^i - I_T^j)(\nabla_{\mathbf{a}} I_T^i - \nabla_{\mathbf{a}} I_T^j)$$

$$\nabla_{\mathbf{a}} \widehat{KP}_T\left[T(\mathbf{a}^t)\right] = -\frac{1}{\sigma_M^2} \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} G_{\sigma_M}(I_T^i, I_T^j)(I_T^i - I_T^j)(\nabla_{\mathbf{a}} I_T^i - \nabla_{\mathbf{a}} I_T^j) .$$

El gradiente puede estimarse con un conjunto de muestreo diferente en cada iteración, introduciendo un factor estocástico al método de ascenso de

gradiente (tal como se propone en [VWI95]), esto permite que el proceso de optimización escape de óptimos locales; en este sentido, el uso del estimador (3.11) es más adecuado debido a que su mayor varianza introduce un componente estocástico adicional. Además de tener el menor costo computacional entre las tres opciones consideradas en este trabajo.

Al trabajar con grandes transformaciones, la parte de la imagen I_R en la región de traslape entre las dos imágenes puede variar con la transformación T, y el gradiente de la similitud debe considerar esta variación. Desafortunadamente no existe una dependencia explícita de I_R con respecto a la transformación; por lo que el gradiente de la similitud debe aproximarse con diferencias finitas. La derivada parcial de (5.6) con respecto a cualquier parámetro a_i puede evaluarse con diferencias finitas centradas como:

$$\frac{\partial \widehat{SKP}}{\partial a_i}\left[T(\mathbf{a}^t)\right] \approx \frac{\widehat{SKP}\left[T(\mathbf{a}^t + \epsilon_i \mathbf{e}_i)\right] - \widehat{SKP}\left[T(\mathbf{a}^t - \epsilon_i \mathbf{e}_i)\right]}{2\epsilon_i}\,, \tag{5.10}$$

en donde $\mathbf{e}_i$ es un vector con un uno en la componente *i-ésima* y ceros en el resto, y ϵ_i es un valor real pequeño. Al emplear esta aproximación la similitud debe evaluarse dos veces por cada parámetro de la transformación y debido a que cada evaluación determina una región de traslape diferente entre las imágenes, para que el gradiente sea calculado con mayor exactitud las muestras empleadas para su estimación deben localizarse en la intersección de todas las regiones de traslape diferentes.

El empleo de (5.10) para la aproximación del gradiente es ventajoso en problemas de registro de imágenes con transformaciones de gran magnitud,

en las que la variación de I_R durante el proceso de registro es importante; de lo contrario puede ignorarse esta variación y emplearse el esquema definido en (5.9). En este trabajo se empleó la aproximación (5.10) para registro de imágenes.

5.4. Registro No-paramétrico.

Para obtener un campo de transformación no-paramétrico (denso), el proceso de registro debe encontrar un vector de traslación diferente para cada punto en las imágenes; en este caso la transformación en cada punto se define de la siguiente manera: $T(x_i) = x_i + u_i$, $i \in \{1, \ldots, N\}$. Para que el campo de transformación no-paramétrico $\mathbf{u} = \{u_1, \ldots, u_N\}$ sea estimado de forma correcta se requiere un gran número de muestras, y el registro por maximización de SKP puede ser prohibitivo debido al costo cuadrático de su evaluación con respecto al tamaño del conjunto de muestreo (por ejemplo, al utilizar el primer estimador, el kérnel se evalúa tomando cada elemento del conjunto de muestreo como primer argumento y cada uno de los elementos restantes en el segundo argumento). En lugar de maximizar la similitud de forma global, ésta puede restringirse a un nivel local, enfocándose en una pequeña región alrededor de cada punto de las imágenes; entonces se puede maximizar la suma sobre todos los puntos x de las similitudes locales. Por ejemplo si se considera una pequeña región cuadrada definida por la ventana W_x, centrada sobre el punto x, entonces la similitud local depende únicamente de los vectores de traslación correspondientes a los puntos contenidos en W_x; estos vectores se representan por el conjunto $\mathbf{v}_x = \{u_i | i \in W_x\}$. Evaluar la similitud a nivel local, además, permite evitar irregularidades en las distribuciones

de las imágenes, las cuales pueden ser resultado de inhomogeneidades espaciales en los valores de intensidad de las imágenes. Finalmente, es necesario considerar la regularización del campo $\mathbf{u}$. Por lo anterior, para registro de no-paramétrico de imágenes, se propone la minimización de la siguiente energía, la cual es una combinación de un término de datos, E_D, y un término de regularización E_S:

$$E(\mathbf{u}) = E_D(\mathbf{u}) + \lambda E_S(\mathbf{u})$$

en donde

$$E_D(\mathbf{u}) = \sum_x \left\{ -\widehat{SKP}_{W_x}(\mathbf{v}_x) \right\} \tag{5.11}$$

$$E_S(\mathbf{u}) = \sum_x \left\{ \sum_{x' \in N_x} \|u_x - u_{x'}\|^2 \right\} \tag{5.12}$$

λ es una constante que regula la suavidad del campo y N_x es un pequeño vecindario alrededor del punto x (independiente de W_x).

La similitud local se evalúa de la siguiente manera:

$$\begin{aligned} \widehat{SKP}_{W_x}(\mathbf{v}_x) &= \frac{\widehat{KP}_J(\mathbf{v}_x)}{\widehat{KP}_T(\mathbf{v}_x)+\widehat{KP}_R(x)} \\ &= \frac{\sum_{i,j\in W_x} K_{\sigma_J}\left(\mathbf{I}_J^i,\mathbf{I}_J^j\right)}{\sum_{i,j\in W_x} K_{\sigma_M}\left(I_T^i,I_T^j\right)+\sum_{i,j\in W_x} K_{\sigma_M}\left(I_R^i,I_R^j\right)} . \end{aligned} \tag{5.13}$$

En este caso $I_T^i = I_S(x_i + u_i)$. Se ha escrito el valor $\widehat{KP}_R$ como una función del punto central x, para remarcar su evaluación local. Se debe notar que se está empleando el estimador (3.12), debido a que al trabajar con

ventanas pequeñas se dispone de pocas muestras para la estimación de las similitudes, y la menor varianza del estimador (3.12) permite un cálculo más exacto del campo de transformación. Resultados similares pueden obtenerse utilizando (3.10), sin embargo debe evitarse el empleo del estimador (3.11), principalmente para ventanas muy pequeñas (ventanas de 3×3 pixeles).

La minimización se realiza mediante descenso de gradiente. Para kérneles gaussianos, (5.7) y (5.8), la derivada parcial del término de datos en la ecuación (5.11) con respecto a un vector de traslación cualquiera u_l es:

$$\frac{\partial E_D}{\partial u_l} = 2 \sum_{x\,:\, l \in W_x} \sum_{i \in W_x} \left\{ \begin{array}{l} f_J(x) G_{\sigma_J}\left(\mathbf{I}_J^l, \mathbf{I}_J^i\right) - \\ f_M(x) G_{\sigma_M}\left(I_T^l, I_T^i\right) \end{array} \right\} \left(I_T^l - I_T^i\right) \nabla I_S(x+u_l) \quad (5.14)$$

donde: $f_J(x) = \frac{1}{\sigma_J^2\left[\widehat{KP}_T(\mathbf{v}_x)+\widehat{KP}_R(x)\right]}$, $f_M(x) = \frac{\widehat{KP}_J(\mathbf{v}_x)}{\sigma_M^2\left[\widehat{KP}_T(\mathbf{v}_x)+\widehat{KP}_R(x)\right]^2}$, y $\nabla I_S(x_l + u_l)$ es el gradiente espacial de la imagen I_S evaluado en el punto $(x_l + u_l)$. Nótese que la primera sumatoria recorre cada una de las ventanas, W_x, que contengan al punto l, mientras que la segunda recorre cada punto contenido en W_x.

Finalmente, el gradiente del término de regularización es

$$\frac{\partial E_S}{\partial u_l} = 4 \left(|N_l| u_l - \sum_{l' \in N_l} u_{l'} \right) .$$

El registro de imágenes por medio de (5.14) puede consumir mucho tiempo al utilizar ventanas de gran magnitud (7×7 pixeles o más). Suponiendo que el valor de KP local ha sido evaluado para un punto x y para un conjunto

de vectores fijo $\mathbf{v}_x^0$, entonces es posible realizar una aproximación del valor de KP para un nuevo conjunto de vectores $\mathbf{v}_x$, empleando una aproximación en serie de Taylor alrededor de $\mathbf{v}_x^0$ en la siguiente forma:

$$\widehat{KP}(\mathbf{v}_x) \approx \widehat{KP}(\mathbf{v}_x^0) + (\mathbf{v}_x - \mathbf{v}_x^0)^T \nabla_{\mathbf{v}} \widehat{KP}(\mathbf{v}_x^0) \ . \tag{5.15}$$

Una vez que los valores de $\widehat{KP}(\mathbf{v}_x^0)$ y $\nabla_{\mathbf{v}}\widehat{KP}(\mathbf{v}_x^0)$ han sido evaluados, el costo de la aproximación del nuevo valor de KP se reduce de $|W|^2$ evaluaciones del kérnel, al cálculo del producto de dos vectores con $|W|$ elementos, sin necesidad de realizar una sola evaluación del kérnel. Sustituyendo las aproximaciones lineales de $\widehat{KP}_J(\mathbf{v}_x)$ y $\widehat{KP}_T(\mathbf{v}_x)$, el valor de (5.13) puede reescribirse como:

$$\widehat{SKP}_{W_x}(\mathbf{v}_x) = \frac{\widehat{KP}_J(\mathbf{v}_x^0) + (\mathbf{v}_x - \mathbf{v}_\mathbf{x}^\mathbf{0})^T \nabla_{\mathbf{v}} \widehat{KP}_J(\mathbf{v}_x^0)}{\widehat{KP}_T(\mathbf{v}_x^0) + (\mathbf{v}_x - \mathbf{v}_x^0)^T \nabla_{\mathbf{v}} \widehat{KP}_T(\mathbf{v}_x^0) + \widehat{KP}_R(x)} \ . \tag{5.16}$$

Con la sustitución de (5.16), el gradiente del término de datos puede simplificarse a:

$$\frac{\partial E_D}{\partial u_l} = - \sum_{x \,:\, l \in W_x} \left\{ f_J(x) \left[\nabla_{\mathbf{v}} \widehat{KP}_J(\mathbf{v}_x^0) \right]_l - f_M(x) \left[\nabla_{\mathbf{v}} \widehat{KP}_T(\mathbf{v}_x^0) \right]_l \right\} \tag{5.17}$$

donde $\left[\nabla_{\mathbf{v}} \widehat{KP}_J(\mathbf{v}_x^0) \right]_l$ y $\left[\nabla_{\mathbf{v}} \widehat{KP}_T(\mathbf{v}_x^0) \right]_l$, son la componente *l-ésima* de los gradientes de KP:

$$\left[\nabla_{\mathbf{v}}\widehat{KP}_M(\mathbf{v}_x^0)\right]_l = -\frac{2}{\sigma_M^2}\sum_{i\in W_x} G_{\sigma_M}\left(I_T^l, I_T^i\right)\left(I_T^l - I_T^i\right)\nabla I_S[x_l + (\mathbf{v}_x^0)_l]$$

$$\left[\nabla_{\mathbf{v}}\widehat{KP}_J(\mathbf{v}_x^0)\right]_l = -\frac{2}{\sigma_J^2}\sum_{i\in W_x} G_{\sigma_J}\left(\mathbf{I}_J^l, \mathbf{I}_J^i\right)(I_T^l - I_T^i)\nabla I_S[x_l + (\mathbf{v}_x^0)_l]$$

y $I_T^i = I_S[x_i + (\mathbf{v}_x^0)_i]$, $f_J(x) = \frac{1}{\widehat{KP}_T(\mathbf{v}_x^0)+(\mathbf{v}_x-\mathbf{v}_x^0)^T\nabla_{\mathbf{v}}\widehat{KP}_T(\mathbf{v}_x^0)+\widehat{KP}_R(x)}$, $f_M(x) = \frac{\widehat{KP}_J(\mathbf{v}_x^0)+(\mathbf{v}_x-\mathbf{v}_x^0)^T\nabla_{\mathbf{v}}\widehat{KP}_J(\mathbf{v}_x^0)}{\left[\widehat{KP}_T(\mathbf{v}_x^0)+(\mathbf{v}_x-\mathbf{v}_x^0)^T\nabla_{\mathbf{v}}\widehat{KP}_T(\mathbf{v}_x^0)+\widehat{KP}_R(x)\right]^2}$.

El registro mediante la minimización de (5.17), requiere de una reevaluación periódica de los valores y gradientes de la kérnel-predictibilidad; en la práctica, después de cada cinco o diez iteraciones. Sin embargo, al emplear la aproximación lineal, puede obtenerse una reducción importante en el tiempo de convergencia sin sacrificar demasiada exactitud.

5.5. Resultados.

Algunos de los resultados obtenidos con la aplicación de SKP en diferentes problemas de registro de imágenes se presentan en esta sección.

5.5.1. Registro Paramétrico.

En el primer conjunto de experimentos se realizó una comparación de los métodos de registro basados en entropía, con respecto al registro por maximización de SKP, utilizando transformaciones afines. Los métodos considerados fueron el registro por maximización de información mutua e información mutua normalizada; para cada uno de ellos se implementaron dos versiones diferentes. La primera implementación se basa en la estimación de la entropía

por medio de histogramas normalizados, y la optimización se realiza con el método simplex [PTVP99]; esta implementación es ampliamente utilizada y sus ventajas sobre otras implementaciones (en todos los casos utilizando histogramas normalizados para estimar la distribución de probabilidad) se describen en [ZC02]. En la segunda implementación, que se basa en el trabajo de Paul Viola [VWI95], la estimación de la entropía se realiza mediante ventanas de Parzen gaussianas [DH73], como se describe a continuación:

$$H(I_R) = -\frac{1}{|A|}\sum_{i\in A}\log\left\{\frac{1}{|B|}\sum_{j\in B}G_{\sigma_M}(I_R^i - I_R^j)\right\} \quad (5.18)$$

$$H[I_L(T)] = -\frac{1}{|A|}\sum_{i\in A}\log\left\{\frac{1}{|B|}\sum_{j\in B}G_{\sigma_M}(I_T^i - I_T^j)\right\} \quad (5.19)$$

$$H[I_L(T), I_R] = -\frac{1}{|A|}\sum_{i\in A}\log\left\{\frac{1}{|B|}\sum_{j\in B}G_{\sigma_J}(I_J^i - I_J^j)\right\}, \quad (5.20)$$

en donde A y B, son dos conjuntos de coordenadas muestreadas en la región de traslape de las imágenes, y G_σ, es la densidad normal con varianza σ^2; el proceso de optimización se realiza utilizando ascenso de gradiente estocástico, aproximando las derivadas parciales por medio de diferencias finitas centradas.

Las transformaciones afines pueden aplicarse multiplicando una matriz cuadrada cualquiera $\mathbf{A}$ por un punto $\mathbf{p}$ y sumando al resultado un vector de traslación $\mathbf{t}$, con lo que se genera el punto transformado $\mathbf{p}'$. La matriz $\mathbf{A}$ está compuesta por tres transformaciones más simples: una rotación $\mathbf{R}$, un escalamiento $\mathbf{S}$ y un cizallamiento $\mathbf{H}$; lo cual se representa en la siguiente expresión:

$$\mathbf{p}' = \mathbf{A}\mathbf{p} + \mathbf{t} = (\mathbf{RSH})\,\mathbf{p} + \mathbf{t}\ .$$

El orden en el cual se multiplican las matrices es arbitrario, y en el caso bidimensional la forma de cada matriz es:

$$\mathbf{R} = \begin{pmatrix} \cos\phi & -\operatorname{sen}\phi \\ \operatorname{sen}\phi & \cos\phi \end{pmatrix}$$

$$\mathbf{S} = \begin{pmatrix} \alpha & 0 \\ 0 & \beta \end{pmatrix}$$

$$\mathbf{H} = \begin{pmatrix} 1 & \gamma \\ 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \delta & 1 \end{pmatrix}\ .$$

Asignando valores aleatorios a los parámetros ϕ, α, β, γ, y δ, se construyeron cinco conjuntos compuestos por 50 transformaciones afines cada uno de ellos. Estos valores aleatorios se generaron muestreando uniformemente intervalos cuya magnitud fue variada, tal como se describe en la tabla 5.1.

Para el registro se utilizaron diferentes imágenes bidimensionales (ver figura 5.1). Las imágenes de referencia se generaron aplicando cambios en intensidad, así como transformaciones afines a las imágenes originales (128×128 pixeles), y extrayendo después un cuadrado de 90×90 pixeles del centro de las imágenes transformadas, tal como se muestra en las figuras, 5.1(a)-5.1(c); lo anterior con excepción de las imágenes 5.1(d) (217×181 pixeles), las cuales

Conjunto	ϕ (grados)	α, β	γ, δ	t (pixeles en cada dirección)
$S1$	$[-10°, 10°]$	$[0.9, 1.1]$	$[-0.1, 0.1]$	$[-10.0, 10.0]$
$S2$	$[-20°, 20°]$	$[0.8, 1.2]$	$[-0.2, 0.2]$	$[-20.0, 20.0]$
$S3$	$[-30°, 30°]$	$[0.7, 1.3]$	$[-0.3, 0.3]$	$[-30.0, 30.0]$
$S4$	$[-40°, 40°]$	$[0.6, 1.4]$	$[-0.4, 0.4]$	$[-40.0, 40.0]$
$S5$	$[-50°, 50°]$	$[0.5, 1.5]$	$[-0.5, 0.5]$	$[-50.0, 50.0]$

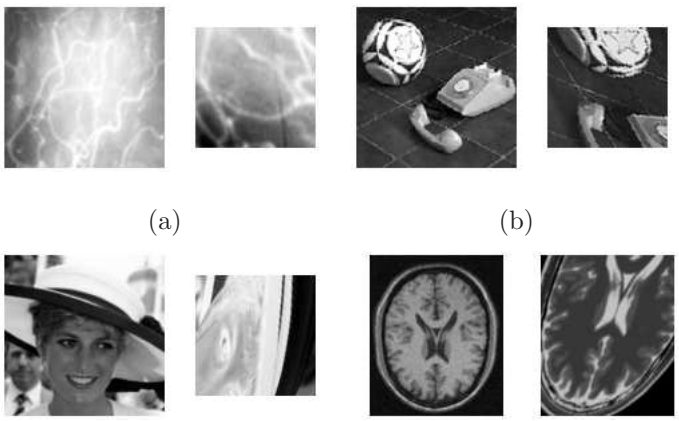
Tab. 5.1: Composición de los cinco conjuntos de transformaciones. El ancho de cada intervalo generador se incrementó progresivamente.

corresponden a dos imágenes de resonancia magnéticas (MRI) obtenidas del simulador del Instituto Neurológico de Montreal [Bra]; en este caso, la imagen I_S corresponde a una MRI de modalidad $T1$ generada con 9 % de ruido y un 40 % de inhomogeneidades en intensidad, mientras que las imágenes de referencia se crearon aplicando transformaciones afines a la correspondiente MRI de modalidad $T2$. Las intensidades de cada par de imágenes fueron escaladas entre 0 y 100; después de eso el cambio en intensidad fue aplicado mediante la función $I_R = 100(\frac{I_L}{100})^{1.35}$ para las imágenes 5.1(a), 5.1(b), e $I_R = 100(1 - \frac{I_L}{100})^{1.35}$ para 5.1(c). Este proceso se repitió para cada transformación de los cinco conjuntos, ejecutando después los algoritmos para registrar los diferentes pares de imágenes.

El registro se realizó bajo un esquema multiescala, empleando dos pirámides gaussianas de tres niveles, las cuales se construyeron aplicando alternativamente las operaciones de suavizado (con un kérnel gaussiano) y submuestreo

a las imágenes I_S e I_R; el registro se inició con la transformación identidad en el nivel más bajo de las pirámides y la transformación obtenida en cada nivel se utilizó como la transformación inicial para los siguientes niveles. En el caso de los algoritmos basados en histogramas normalizados los detalles de implementación se fijaron de acuerdo a lo sugerido por [ZC02]. En el caso de los algoritmos basados en ventanas de Parzen, se utilizaron dos conjuntos de coordenadas diferentes compuestos de 50 muestras cada uno. En la estimación de la entropía conjunta, la matriz de covarianza se tomó igual a un múltiplo de la matriz identidad $\sigma^2 I$, mientras que para las entropías marginales la varianza se fijó en el valor σ^2; este valor se estableció manualmente, considerando un porcentaje del rango dinámico de las imágenes a registrar. Los valores utilizados en estos experimentos fueron $\sigma = 5\,\%$ para 5.1(a) y $\sigma = 10\,\%$ en el resto de las imágenes. En el caso de registro por maximización de SKP, se utilizó el estimador (3.11) con el mismo número de muestras que se emplearon en los algoritmos basados en ventanas de Parzen; además de que el ancho de los kérneles gaussianos se seleccionó bajo el mismo criterio descrito anteriormente, excepto que en todos los experimentos se fijó $\sigma = 8\,\%$.

El número de registros realizados correctamente en cada conjunto y para cada algoritmo se muestra en la figura 5.2. Para esto, el registro de dos imágenes se consideró correcto si el error promedio entre el campo aplicado y el recuperado fue menor a un pixel. Puede observarse que, casi en todos los casos, el número de registros realizados correctamente mediante maximización de SKP fue superior que los obtenidos con el resto de los métodos



(a) (b)

(c) (d)

Fig. 5.1: Imágenes empleadas en el registro. Las imágenes de referencia para 5.1(a)-5.1(c) se obtuvieron aplicando cambios de intensidad y transformaciones afines a las imágenes originales, y extrayendo después un pequeño cuadrado del centro de la imagen transformada. Para 5.1(d), las imágenes de referencia fueron obtenidas aplicando transformaciones afines a una imagen de resonancia magnética de modalidad $T2$.

analizados, especialmente en transformaciones de gran magnitud; además se observa muy poca robustez de los algoritmos de registro basados en histogramas normalizados. La precisión de los métodos de registro analizados está relacionada directamente con el método de optimización empleado, observándose una mayor precisión en los métodos de registro basados en el método simplex downhill. En el caso de los métodos optimizados por descenso de gradiente estocástico con gradiente aproximado con diferencias finitas, se obtuvo una menor precisón, esto posiblemente al ruido introducido en la evaluación del gradiente.

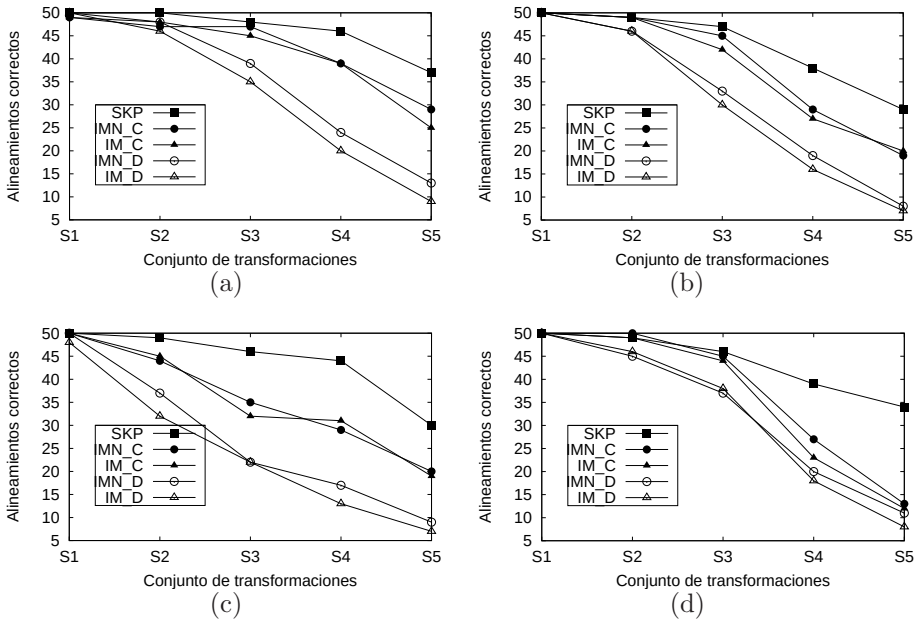


Fig. 5.2: Registros realizados correctamente en función de la complejidad de la transformación. La gráfica muestra los resultados correspondientes a las imágenes 5.1(a)-5.1(d). En la gráfica, *SKP*, representa la similitud basada en kérnel-predictibilidad, *NMI-D* y *NMI-D* identifican a la información mutua normalizada basada en ventanas de Parzen e histogramas normalizados respectivamente; mientras que *MI-C* y *MI-D*, identifican a la información mutua basada en ventanas de Parzen e histogramas normalizados.

El registro por maximización de SKP presenta otra ventaja al compararse con los algoritmos basados en la estimación de la entropía mediante ventanas de Parzen. Debido al costo cuadrático de la estimación tanto de la entropía como de la KP, el número de muestras utilizadas para realizar el registro es un parámetro muy importante. La figura 5.3 muestra el desempeño de los tres métodos al variar este parámetro; en este caso se empleó el conjunto de transformaciones afines $S3$ (descrito en la tabla 5.1) en las cuatro imágenes. Puede observarse que el registro realizado por maximización de SKP puede realizarse aceptablemente incluso al emplear un conjunto de muestras reducido; a diferencia del registro por IM e IMN. En algunas de las gráficas se muestra una pequeña reducción en el desempeño del registro por SKP al utilizar conjuntos de muestreo de gran tamaño, este comportamiento está causado por la disminución en el ruido introducido en la aproximación del gradiente de la SKP, ya al aumentar la cardinalidad del conjunto de muestreo se disminuye la varianza de los estimadores utilizados, haciendo que el registro sea un poco más sensible a la presencia de óptimos locales.

Con el fin de evaluar el registro por maximización de SKP al emplear diferentes kérneles, se repitió el proceso de registro de los cuatro pares de imágenes (mostrados en la figura 5.1) para SKP, utilizando los kérneles unidimensionales descritos en la tabla 5.5.1. Estos kérneles se emplearon en la evaluación de los valores de SKP marginales mientras que para el cálculo de la KP conjunta se empleó un kérnel separable de acuerdo a la ecuación (5.2). Como puede observarse en las figuras 5.4(a)-5.4(d), la selección del kérnel para el registro por maximización de SKP no es un factor crítico, ya

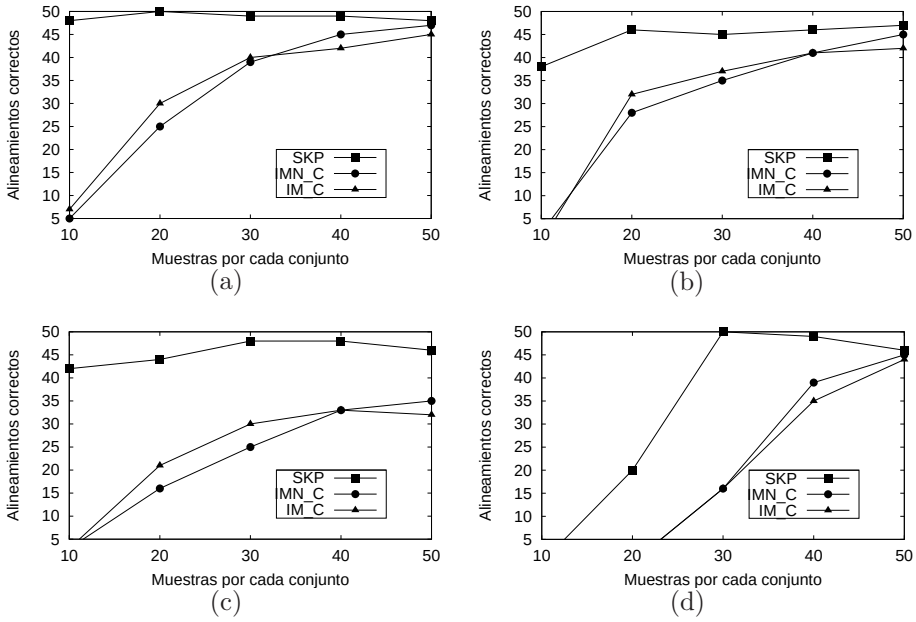


Fig. 5.3: Registros realizados correctamente en función del número de muestras utilizados para la estimación de las medidas. La gráfica muestra los resultados correspondientes a las imágenes 5.1(a)-5.1(d).

Kérnel gaussiano	$K(x_1, x_2) = \exp\left[-(x_1 - x_2)^2/\sigma^2\right]$
Kérnel Cauchy	$K(x_1, x_2) = \frac{1}{1+\alpha(x_1-x_2)^2}$
Kérnel exponencial	$K(x_1, x_2) = \exp\left(-\lvert x_1 - x_2\rvert/\sigma^2\right)$
Kérnel triangular	$K(x_1, x_2) = 1 - \alpha\lvert x_1 - x_2\rvert$ para $\alpha\lvert x_1 - x_2\rvert < 1$ y $K(x_1, x_2) = 0$, en otro caso.

Tab. 5.2: Diferentes kérneles utilizados en el registro por SKP.

que solamente se obtuvieron pequeñas diferencias en desempeño al utilizar diferentes kérneles suaves, sin embargo, los resultados obtenidos fueron muy pobres en el caso del kérnel triangular.

Finalmente, en la figura 5.5 se muestra una comparación del desempeño de las dos medidas de similitud que se describieron en la sección 5.1. Dado que las diferencias son muy pequeñas, el costo computacional se convierte en el factor que justifica la selección de la medida de similitud 5.1.

5.5.2. Registro No-paramétrico

La robustez obtenida bajo transformaciones de gran magnitud y empleando un número reducido de muestras, hacen que el registro por maximización de SKP sea adecuado para aplicarse en problemas de registro no-paramétrico. Para evaluar el desempeño de $\widehat{SKP}$ en este tipo de problemas, se generaron sinteticamente diez campos vectoriales mediante dos mallas de 15×15 nodos, en los cuales se centraron funciones B-spline cúbicas y se asignaron valores aleatorios a cada nodo; de manera que para cada pixel (x, y) se definió un vector de traslación $(u(x, y), v(x, y))$ de la siguiente manera:

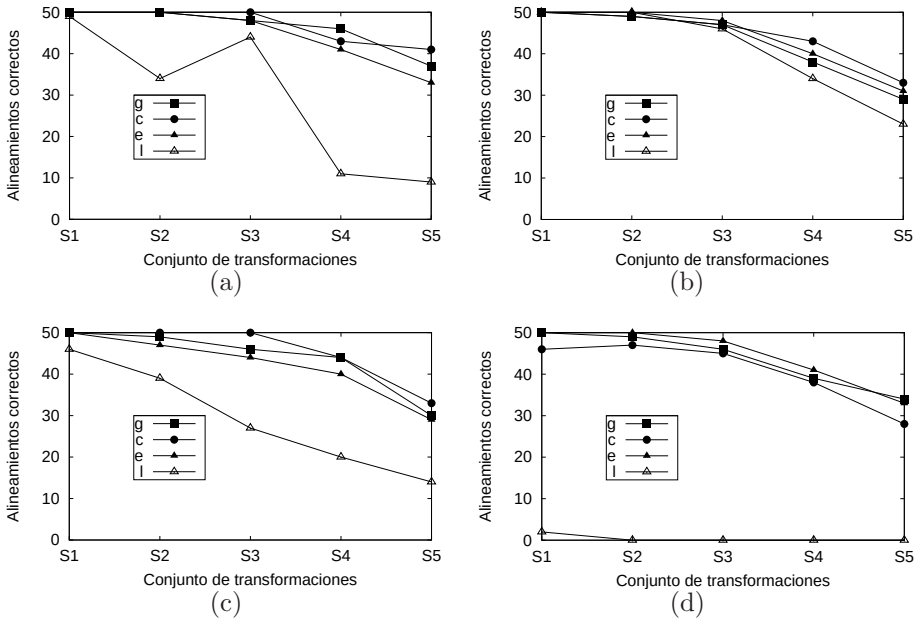


Fig. 5.4: Resultados del registro mediante SKP utilizando diferentes kérneles. La letra g identifica al kérnel gaussiano, la letra c al de Cauchy, la e al exponencial y la letra l al kérnel triangular.

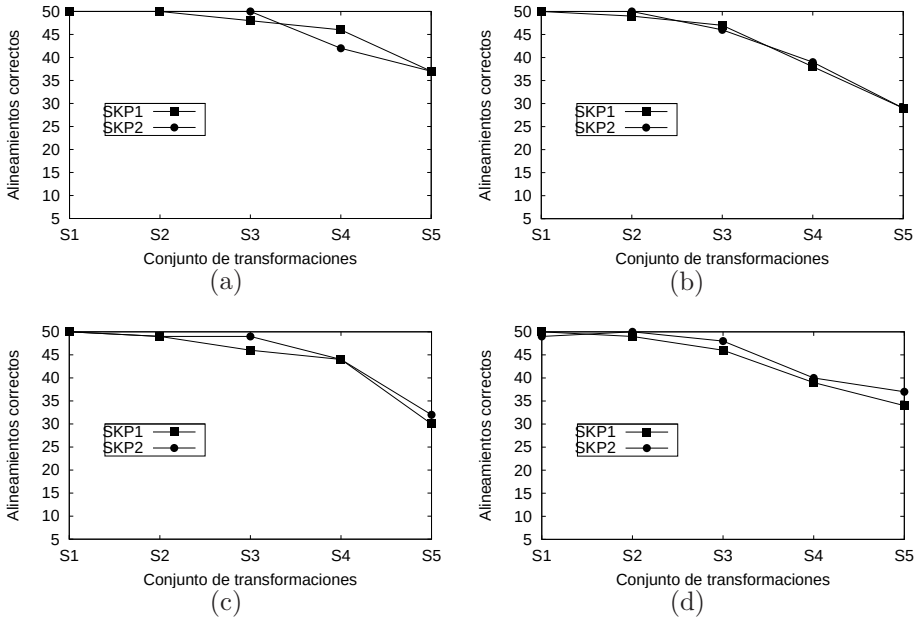


Fig. 5.5: Comparación de los resultados obtenidos en la utilización de las dos medidas de similitud basadas en SKP (denominadas SKP_1 y SKP_2, ver ecuaciones 5.1 y 5.5).

$$\begin{aligned} u(x,y) &= \sum_{i=1}^{15}\sum_{j=1}^{15} U_{ij}\beta[k_1(x-x_i)]\beta[k_2(y-y_j)] \\ v(x,y) &= \sum_{i=1}^{15}\sum_{j=1}^{15} V_{ij}\beta[k_1(x-x_i)]\beta[k_2(y-y_j)] \end{aligned} \tag{5.21}$$

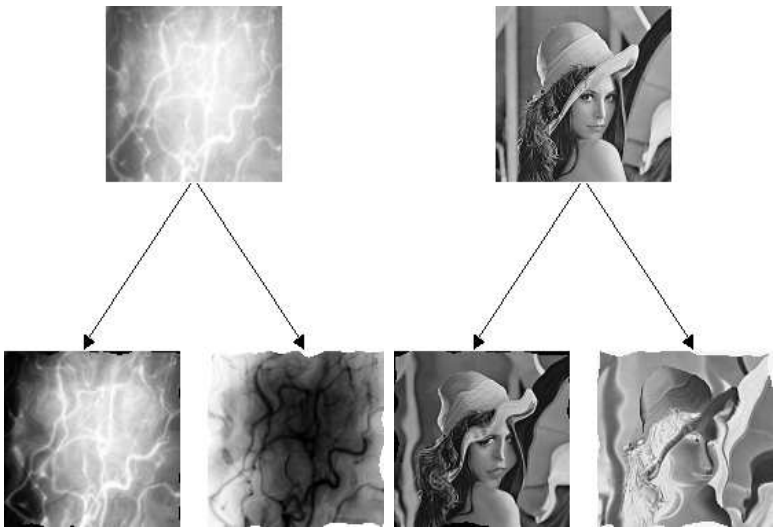
donde $U_{ij}, V_{ij} \sim U\{-7,7\}$, en todos los nodos (x_i, y_j), y k_d es la razón de nodos sobre la magnitud en pixeles para la dirección d. Las funciones B-spline utilizadas son las siguientes:

$$\beta(z) = \begin{cases} \frac{2}{3} - |z|^2 + \frac{|z|^3}{2}, & |z| < 1. \\ \frac{(2-|z|)^3}{6}, & 1 \leq |z| < 2 \\ 0, & |z| \geq 2. \end{cases}$$

Los campos vectoriales generados se aplicaron a dos imágenes diferentes después de cambiar los tonos de gris mediante las funciones $f_1(I) = 100(\frac{I}{100})^{1.35}$ y $f_2(I) = 100(1 - \frac{I}{100})^{1.35}$, como se muestra en la figura 5.6. Después el algoritmo de registro no-paramétrico descrito en la sección 5.4 fue ejecutado buscando recuperar el campo vectorial originalmente aplicado. En cada caso se evaluó el error, definido como el promedio de las longitudes de los vectores resultantes de la diferencia entre el campo aplicado y el obtenido por el registro. Al igual que en el caso paramétrico, el registro se realizó utilizando una representación multiescala de las imágenes mediante pirámides gaussianas de tres niveles, iniciando con la transformación identidad en el nivel más bajo y en el resto de los niveles con el campo vectorial resultante del registro en el nivel previo. Para comparar, el algoritmo de registro se ejecutó sustituyendo la $\widehat{SKP}$ en el término (5.11) por expresiones correspondientes a IM e IMN basadas en ventanas de Parzen: ecuaciones (5.18)-(5.20). Como se describe en la sección 5.4, las medidas de similitud se evaluaron a nivel local, medi-

ante pequeñas ventanas centradas sobre cada pixel. El desempeño de las tres medidas de similitud se evaluó al utilizar ventanas de diferente tamaño. Los resultados obtenidos se resumen en las figuras 5.7(a)-5.7(d), en las que se observa que pueden obtenerse importantes reducciones en el error promedio al emplear SKP como medida de similitud, sobre todo al realizar el registro con ventanas pequenãs. Lo anterior se refleja en un ahorro importante en el tiempo de registro (ver figura 5.8). Para facilitar una comparación cualitativa de los errores, las imágenes registradas utilizando las tres medidas de similitud en una transformación específica se muestran en la figura 5.9.

La utilización de la aproximación lineal del valor de SKP (como se describe en la sección 5.4) permite acelerar la convergencia del registro sin sacrificar demasiada exactitud. En la figura 5.5.2 se muestra una comparación de la convergencia del error con respecto al tiempo de ejecución, tanto del algoritmo de registro original basado en SKP como del basado en la aproximación lineal. Ambos algoritmos se ejecutaron sobre el mismo par de imágenes y el registro se basó en ventanas de 5×5 pixeles.



(a) (b)

Fig. 5.6: Imágenes utilizadas en el registro no-paramétrico. Las imágenes de referencia fueron generadas aplicando cambios en la intensidad así como diferentes campos vectoriales a las imágenes originales.

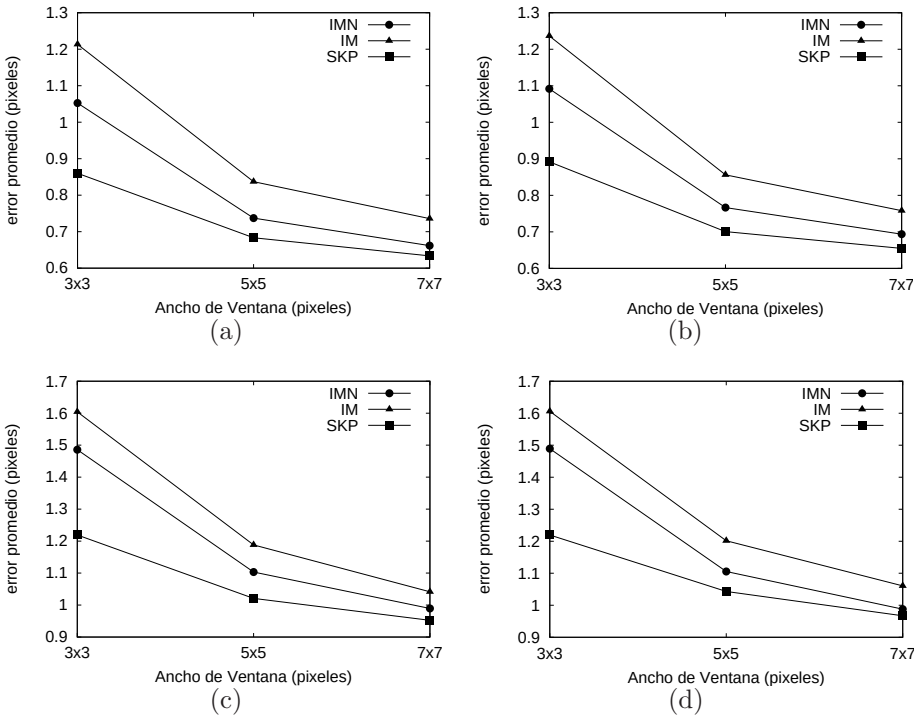


Fig. 5.7: Error promedio en el registro no-paramétrico para diferentes tamaños de ventana. El primer renglón muestra los resultados obtendios en el registro de la imagen 5.6(a) e imágenes de referencia generadas mediante la función de transferencia de tonos $f_1(I) = 100(\frac{I}{100})^{1.35}$ (gráfica de la izquierda), y $f_2(I) = 100(1 - \frac{I}{100})^{1.35}$ (derecha). La segunda fila muestra los resultados correspondientes a la imagen 5.6(b).

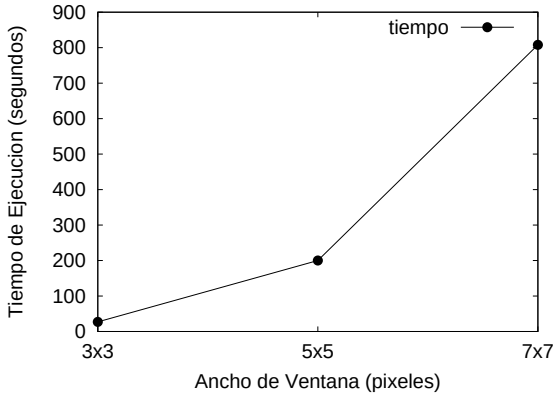


Fig. 5.8: Tiempo de ejecución del registro no-paramétrico con SKP como una función del ancho de las ventanas utilizadas para medir la similitud local. Se muestran los resultados obtenidos para una imagen de 128×128 pixeles. Para cada ventana se realizaron 200 iteraciones del algoritmo de descenso de gradiente en cada nivel de la pirámide gaussiana. Las pruebas fueron ejecutadas en una PC con procesador Pentium 4, a 3.0 GHz.

(a) (b) (c)

Fig. 5.9: Imágenes registradas para una transformación específica. Se muestra la imagen de referencia en la primera fila (la misma en cada caso), y en la segunda las imágenes registradas por SKP (izquierda), NMI (centro) y MI (derecha). La estimación del campo de deformación se realizó utilizando ventanas de 3×3 pixeles alrededor de cada pixel en las imágenes. Los errores respectivos fueron de: 1.23, 1.57 and 1.60 pixeles.

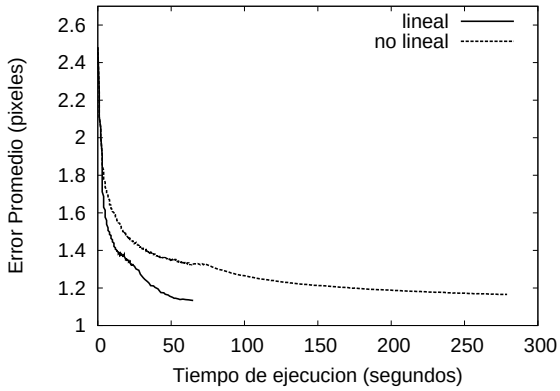


Fig. 5.10: Convergencia para los algoritmos de registro no-paramétrico basados en SKP. Se muestran los resultados obtenidos al registrar un mismo par de imágenes tanto con el algoritmo original como con la aproximación lineal.

6. SEGMENTACIÓN DE IMÁGENES MEDIANTE LA MAXIMIZACIÓN DE LA KÉRNEL PREDICTIBILIDAD REGIONAL.

La segmentación de una imagen consiste en la asignación de una etiqueta, perteneciente a un conjunto discreto $L = \{l_1, l_2, \ldots, l_K\}$, a cada uno de sus puntos. A través de las etiquetas es posible la identificación de patrones regulares en imágenes como pueden ser tonos de gris, texturas, estructuras anatómicas en imágenes médicas, valores de disparidades en secuencias, entre otros. Estos patrones quedan determinados por la formación de la imagen observada, g, a través del siguiente proceso:

$$g(x) = \Phi(x, N_x, m_k) \oplus R_x, \ \forall x \in \{1, \ldots, N\} \ , \tag{6.1}$$

siendo $g(x)$ el valor de la intensidad de la imagen obsevada en el pixel x.

En este proceso Φ es una función de transferencia de tonos, mientras que N_x representa una vecindad del punto x. Asociado a cada patrón (o modelo) se tiene un conjunto de parámetros m_k, $k \in \{1, \ldots, K\}$. Finalmente el proceso de formación se ve afectado por la presencia de ruido, R_x, aplicado a través de alguna operación invertible, $\oplus$.

Entre los métodos tradicionalmente utilizados para segmentar imágenes, algunos de los más destacados son los basados en crecimiento y unión de regiones [YC08], los basados en curvas dinámicas [ZY96, CV01, PD99], y los basados en la teoría bayesiana para la toma de decisiones. Estos últimos, gozan de una gran aceptación debido a la generalidad de sus aplicaciones y serán descritos en la siguiente sección.

6.1. Segmentación de Imágenes como un Problema de Toma de Decisiones.

La segmentación de una imagen es equivalente a un problema de toma de decisiones sobre el valor del campo aleatorio de etiquetas $\mathbf{F} = \{F_1, F_2, \ldots, F_N\}$ con $F_i \in L$, $\forall i$. Bajo este enfoque, la segmentación de la imagen se basa en la asignación de una función de costo, $C(\mathbf{f}, \mathbf{f}^*)$, a la decisión de etiquetar la imagen con el campo $\mathbf{f}$ (haciendo $\mathbf{F} = \mathbf{f}$), cuando el verdadero valor es $\mathbf{f}^*$. Dado que $\mathbf{f}^*$ es desconocido, se selecciona el campo $\widehat{\mathbf{f}}$ que minimice el valor esperado de la función de costo:

$$\widehat{\mathbf{f}} = \arg\min_{\mathbf{f}} \sum_{\mathbf{f}^*} C(\mathbf{f}, \mathbf{f}^*) P(\mathbf{F} = \mathbf{f}^* | g) \tag{6.2}$$

en donde $P(\mathbf{F} = \mathbf{f}^*|g)$ es la distribución *a posteriori*, la cual puede obtenerse mediante el teorema de Bayes:

$$P(\mathbf{F} = \mathbf{f}^*|g) = \frac{P(g|\mathbf{F} = \mathbf{f}^*)P(\mathbf{F} = \mathbf{f}^*)}{P(g)} \, . \tag{6.3}$$

La probabilidad $P(g)$ puede ignorarse ya que su valor es constante, mientras que, considerando independencia espacial en el ruido que afecta a la formación de la imagen observada, la probabilidad $P(g|\mathbf{F}=\mathbf{f}^*)$, denominada *verosimilitud*, es equivalente igual a:

$$P(g|\mathbf{F}=\mathbf{f}^*) = \prod_{x=1}^{N} P\left\{R_x = \left[g(x) \ominus \Phi(x, N_x, m_{f_x^*})\right]\right\} \tag{6.4}$$

siendo $\ominus$ la operación inversa de $\oplus$, y $m_{f_x^*}$ los parámetros del modelo asociado con la etiqueta f_x^*.

Mención especial merece el término $P(\mathbf{F}=\mathbf{f}^*)$, denominado distribución de probabilidad *a priori*. A través de este término es posible reflejar ciertos criterios acerca de las características del campo de etiquetas que pueden ser deseables, por ejemplo condiciones de regularidad, lo cual guía fuertemente el proceso de segmentación. Esta distribución es frecuentemente representada a través de *campos aleatorios markovianos* los cuales se describen en la siguiente sección.

Existen varias alternativas para la selección de la función de costo. Cuando el conjunto de etiquetas es ordenado, entonces una posibilidad consiste en hacer $C(\mathbf{f}, \mathbf{f}^*) = \|\mathbf{f} - \mathbf{f}^*\|^2$; con lo que el estimador obtenido en (6.2) es igual a la media de la distribución a posteriori:

$$\widehat{\mathbf{f}} = \sum_{\mathbf{f}} \mathbf{f} P(\mathbf{F}=\mathbf{f}|g) \ .$$

Mientras que una opción ampliamente utilizada se obtiene haciendo $C(\mathbf{f}, \mathbf{f}^*) = 1 - \delta(\mathbf{f} - \mathbf{f}^*)$ (la función de costo que asigna un uno si $\mathbf{f} \neq \mathbf{f}^*$ y cero en caso

contrario). En este caso el estimador de (6.2) se reduce a:

$$\widehat{\mathbf{f}} = \arg \max_{\mathbf{f}} P(\mathbf{F} = \mathbf{f}|g) \ ,$$

también conocido como estimador *máximo a posteriori* (MAP).

La función que da origen al estimador MAP asigna costos constantes a campos de etiquetas que son diferentes del verdadero valor, independientemente del número de puntos en que puedan llegar a coincidir. Otra opción consiste en seleccionar la función de costo igual al número de puntos en los que ambos campos difieren, con lo que el estimador óptimo es el maximizador de las *marginales a posteriori* [MMP87]:

$$\widehat{f_i} = \arg \max_{k} P(F_i = l_k|g) = \arg \max_{k} \sum_{\mathbf{f}:f_i=l_k} P(\mathbf{f}|g), \ \forall i \ .$$

Independientemente del estimador utilizado, debe notarse que el problema planteado en (6.2) es combinatorio, y que el espacio de soluciones tiene una cardinalidad igual a K^N, por lo que la optimización llega a ser un problema crítico.

6.2. Campos Aleatorios Markovianos y Gibbsianos.

Dado un conjunto de sitios, $S = \{s_1, s_2, \ldots, s_N\}$, un sistema de vecindades, $N_i \subset S$, definido en cada sitio s_i, y un campo de variables aleatorias, $\mathbf{F} = \{F_1, F_2, \ldots, F_N\}$; se dice que $\mathbf{F}$ es un *campo aleatorio markoviano* si satisface las siguientes condiciones [Li01]:

$$P(\mathbf{F} = \mathbf{f}) \geq 0 \,, \tag{6.5}$$

$$P(F_i = f_i | F_j = f_j, j \neq i) = P(F_i = f_i | F_j = f_j, j \in N_i) \,, \forall i \,. \tag{6.6}$$

La propiedad de markovianidad en un campo aleatorio refleja características que dependen de interacciones locales entre las variables, dado que se restringe la dependencia de cualquier variable F_i, únicamente a las variables definidas en sitios ubicados en la vecindad del sitio s_i. Esta dependencia puede dificultar, a primera vista, la construcción de la distribución conjunta, $P(\mathbf{F} = \mathbf{f})$; sin embargo, puede aprovecharse la equivalencia entre un campo aleatorio markoviano y un *campo aleatorio gibbsiano*, establecida mediante el teorema de Hammersley-Clifford [Li01], para superar esta dificultad.

Un campo aleatorio gibbsiano, definido sobre el conjunto de sitios S, tiene una función de probabilidad igual a:

$$P(\mathbf{F} = \mathbf{f}) = \frac{1}{Z} \exp\left(-\frac{U(\mathbf{f})}{T}\right) \,, \tag{6.7}$$

siendo $Z = \sum_{\mathbf{f}} \exp\left(-\frac{U(\mathbf{f})}{T}\right)$ y T una constante que controla la anchura de la distribución. La función, $U(\mathbf{f})$, es una función de energía que debe poder representarse mediante una expresión con la siguiente forma:

$$U(\mathbf{f}) = \sum_{c} V_c(\mathbf{f}) \,, \tag{6.8}$$

en donde la sumatoria se extiende por todos los *cliques*, c, del conjunto de sitios. Un clique, se define como un subconjunto de S tal que cada par de

sitios diferentes contenidos en él son vecinos [GG84]. Finalmente, la función V_c es conocida como potencial.

El camino a seguir para construir la distribución $P(\mathbf{F} = \mathbf{f})$, cuando $\mathbf{F}$ es un campo aleatorio markoviano, inicia con la selección de un potencial que refleje adecuadamente las características locales del campo sobre un sistema de vecindades definido. Después de evaluar la energía (6.8), ésta se sustituye en (6.7) dada la igualdad entre el campo markoviano y uno gibbsiano.

La construcción explícita de $P(\mathbf{F} = \mathbf{f})$ llega a ser prohibitiva debido a la necesidad de evaluar la constante de normalización Z, sin embargo, en algunas aplicaciones esto es innecesario, por ejemplo, al utilizar el estimador MAP y cuando los valores de los parámetros, m_k, son conocidos para todo valor de k. Bajo estas circunstancias, la distribución a posteriori (6.3) puede escribirse como una distribución gibbsiana:

$$P(\mathbf{F} = \mathbf{f}|g) = \frac{1}{Z} \exp\left(-\frac{U'(\mathbf{f})}{T}\right) ,$$

bajo una nueva función de energía igual a:

$$U'(\mathbf{f}) = -\sum_i \ln v_{f_i} + \sum_c V_c(\mathbf{f}) , \tag{6.9}$$

siendo $v_{f_i} = P(F_i = f_i|g)$. Con estas suposiciones, la segmentación por estimador MAP es equivalente a la búsqueda del valor $\widehat{\mathbf{f}}$ tal que:

$$\widehat{\mathbf{f}} = \arg\min_{\mathbf{f}} U'(\mathbf{f}) .$$

Aún utilizando (6.2), la segmentación sigue representando un problema de optimización combinatoria. Algunas metodologías que derivan en funciones de energía con mejores condiciones de optimización se analizan en la siguiente sección.

6.3. Campos de Medida Aleatorios Markovianos Ocultos.

Puede agregarse un paso intermedio en la generación de la imagen observada g, al suponer que el campo de etiquetas $\mathbf{f} = \{f_1, f_2, \ldots, f_N\}$ es una realización de un campo aleatorio markoviano oculto $\mathbf{p}$, [MSB03]. Este campo, asigna sobre cada sitio un vector de probabilidades continuas:

$$p(s_i) = \{p_1(s_i), p_2(s_i) \ldots, p_K(s_i)\}$$

con las restricciones:

$$\sum_k p_k(s_i) = 1, \ \forall i, \tag{6.10}$$

$$p_k(s_i) \geq 0, \ \forall i, k \ , \tag{6.11}$$

las cuales establecen que cada vector de probabilidades debe pertenecer al *simplejo* K-dimensional.

Suponiendo independencia espacial, la distribución del campo $\mathbf{F}$ condicionado a $\mathbf{p}$ es igual a:

$$P(\mathbf{F} = \mathbf{f}|\mathbf{p}) = \prod_{s_i \in S} p_{f_i}(s_i) \ . \tag{6.12}$$

Una vez conocido el valor de $\mathbf{p}$, segmentar la imagen consiste en seleccionar el valor $\mathbf{f}$ que maximice (6.12), lo cual es equivalente a tomar:

$$\widehat{f_i} = \arg \max_k p_k(s_i) \ .$$

El valor del campo $\mathbf{p}$ puede estimarse maximizando la distribución a posteriori:

$$P(\mathbf{p}|g) = \frac{P(g|\mathbf{p})P(\mathbf{p})}{P(g)} \ ,$$

y la verosimilitud de la imagen observada con respecto al campo markoviano, se obtiene usando como base la expresión:

$$P(g(s_i), f_i|\mathbf{p}) = P(g(s_i)|f_i, \mathbf{p})P(f_i|\mathbf{p}) \ ,$$

marginalizando y considerando que $P(f_i|\mathbf{p}) = p_{f_i}(s_i)$ y además que $P(g(s_i)|f_i, \mathbf{p}) = P(g(s_i)|f_i) = v_{f_i}(s_i)$:

$$\begin{aligned} P\left[g(s_i)|\mathbf{p}\right] &= \sum_k P\left[g(s_i), k|\mathbf{p}\right] \\ &= \sum_k v_k(s_i)p_k(s_i) \end{aligned} \tag{6.13}$$

La energía gibbsiana correspondiente es:

$$U(\mathbf{p}) = -\sum_i \ln \left\{ \sum_k v_k(s_i)p_k(s_i) \right\} + \sum_c V_c(\mathbf{p}) \ .$$

Dado que el campo $\mathbf{p}$ es continuo, la minimización de la energía gibbsiana puede realizarse a través de algún método de gradiente, lo cual representa una gran ventaja. Únicamente debe tenerse cuidado en respetar las retricciones (6.10) y (6.11); en la práctica los valores del campo markoviano se restringen realizando una proyección al simplejo cada vez que se violan las restricciones.

Cuando el vector de probabilidades, $p(s_i)$, tiene una baja entropía (es muy cercano a un vector de probabilidades binario), la energía gibbsiana (6.3) puede aproximarse por la siguiente expresión:

$$U(\mathbf{p}) = -\sum_i \left\{ \sum_k \ln\left[v_k(s_i)\right] p_k^2(s_i) \right\} + \sum_c V_c(\mathbf{p}) \ . \tag{6.14}$$

Más aún, la condición de baja entropía puede forzarse en la energía gibbsiana minimizando la entropía de Gini de cada vector de probabilidades, lo cual da origen al modelo de segmentación basado en *campos de medida gaussianos-markovianos con entropía controlada* [ROM05, ROM07].

$$U(\mathbf{p}) = -\sum_i \left\{ \sum_k p_k^2(s_i) \left[\ln v_k(s_i) - \mu\right] \right\} + \sum_c V_c(\mathbf{p}) \ , \tag{6.15}$$

en donde μ es una constante que controla la restricción de mínima entropía en la distribución $p(s_i)$.

6.4. Kernel-Predictibilidad Regional como Conocimiento a Priori.

Indudablemente la opción más utilizada para representar la distribución a priori es la de imponer restricciones de suavidad sobre el campo de eti-

quetas. Esta condición asigna valores de probabilidad más altos a campos que presentan pocas diferencias entre sitios vecinos. En el caso de etiquetas ordenadas, por ejemplo, las condiciones de suavidad pueden establecerse mediante la siguiente distribución:

$$P(\mathbf{F} = \mathbf{f}) = \exp\left[-\sum_{i}\sum_{j\in N_i}\left(f_{s_i} - f_{s_j}\right)^2\right], \tag{6.16}$$

la cual asigna una energía gibbsiana que penaliza una aproximación en diferencias finitas a la magnitud cuadrada del gradiente del campo de etiquetas. En muchas aplicaciones, sin embargo, este potencial tiene el inconveniente de sobre-difundir los modelos en regiones en donde el problema de segmentación está pobremente determinado, por ejemplo en regiones sin textura dentro de aplicaciones relacionadas con movimiento.

En muchas aplicaciones, parte del conocimiento a priori puede establecerse mediante el criterio de realizar una segmentación utilizando la menor cantidad de modelos para describir los patrones encontrados en cada objeto, dentro de una escena. Esta hipótesis es plausible en problemas como la segmentación de disparidades estereoscópicas y de movimiento, por ejemplificar, siempre y cuando los modelos permitan suficientes grados de libertad. La información para identificar objetos puede extraerse de una simple segmentación de tonos de gris o de color. De manera que, suponiendo que se tiene disponible una segmentación previa de la imagen, identificada como I_S, se busca que la nueva segmentación describa con el menor número de modelos las regiones homogéneas en I_S. Sea R_i una de estas regiones, el criterio descrito anteriormente puede formalizarse buscando minimizar alguna medida

de información sobre la distribución de probabilidad de las nuevas etiquetas en cada R_i. Si $P(F_{R_i} = k)$ representa la probabilidad de que el campo de etiquetas tenga el valor k en la región R_i, la cual puede evaluarse mediante la siguiente expresión:

$$P(F_{R_i} = k) = \frac{1}{|R_i|} \sum_{s \in R_i} \delta(F_s - k) \ , \tag{6.17}$$

entonces la medida de información puede ser el negativo de la kérnel-predictibilidad; con lo que se propone la siguiente energía gibbsiana:

$$U(\mathbf{f}) = -\sum_i W\left(|R_i|\right) \left[\sum_k \sum_{k'} P(f_{R_i} = k) K(k, k') P(f_{R_i} = k') \right] \ , \tag{6.18}$$

la cual es equivalente a una suma ponderada de la kérnel-predictibilidad de cada región R_i. La función $W\left(|R_i|\right)$ permite asignar mayor o menor importancia en la energía a algunas regiones, de acuerdo a sus características.

La energía definida en (6.18), hace que el campo de etiquetas $\mathbf{F}$ sea markoviano si se define la vecindad de cada sitio, $s_k \in R_i$, como el resto de los sitios en R_i. Más aún, la clásica restricción de suavidad puede complementar este criterio, por lo que se propone la siguiente energía gibbsiana para construir la distribución a priori:

$$\begin{aligned} U(\mathbf{f}) \ = \ & \sum_{s_i} \sum_{s_j \in N_i} V(f_{s_i}, f_{s_j}) - \\ & \lambda \sum_i W\left(|R_i|\right) \left[\sum_k \sum_{k'} P(f_{R_i} = k) K(k, k') P(f_{R_i} = k') \right] \ , \end{aligned} \tag{6.19}$$

en donde se aplica el potencial de suavidad V a s_i y a cada sitio ubicado en una vecindad adecuada de s_i, N_i. La vecindad del sitio que define el nuevo campo markoviano estará determinada por la unión de N_i y R_m (exceptuando a s_i), siendo R_m la región homogénea a la cual pertenece s_i. Finalmente, la constante positiva λ permite controlar el peso de la kérnel predictibilidad en la energía. Sustituyendo la expresión (6.17) en la energía (6.19), ésta puede reducirse a:

$$\begin{aligned} U(\mathbf{f}) &= \sum_{s_i} \sum_{s_j \in N_i} V(f_{s_i}, f_{s_j}) - \\ &\quad \lambda \sum_i W'(|R_i|) \left[\sum_{s \in R_i} \sum_{s' \in R_i} K(F_s, F_{s'}) \right] , \end{aligned} \tag{6.20}$$

siendo $W'(|R_i|) = \frac{W(|R_i|)}{|R_i|^2}$.

De manera particular, y para simplificar, puede elegirse el kérnel igual a la delta de Kronecker, con lo que se obtiene la siguiente expresión para la energía gibbsiana:

$$\begin{aligned} U(\mathbf{f}) &= \sum_{s_i} \sum_{s_j \in N_i} V(f_{s_i}, f_{s_j}) - \lambda \sum_i W(|R_i|) \left[\sum_k P^2(f_{R_i} = k) \right] \\ &= \sum_{s_i} \sum_{s_j \in N_i} V(f_{s_i}, f_{s_j}) - \lambda \sum_i W'(|R_i|) \left[\sum_{s \in R_i} \sum_{s' \in R_i} \delta(f_s - f'_s) \right] \end{aligned} \tag{6,21}$$

el segundo término es proporcional a la suma de la entropía de Gini de la distribución de etiquetas sobre cada región.

Esta propuesta debe contrastarse con otros enfoques explorados anteriormente y que también intentan aprovechar la información proveniente de segmentaciones previas de la escena en tonos de gris o de color. Uno de los más comunes consiste en la utilización de un término de suavidad no homogéneo, pues se reduce la penalización en las fronteras de las regiones R_i, con lo que se busca que las transiciones fuertes en el campo de etiquetas coincidan con las interfases de la segmentación de intensidad. Este enfoque, sin embargo, puede resultar contraproducente en regiones finamente texturizadas, en las que existe una gran densidad de fronteras debido a que las regiones homogéneas son pequeñas; bajo esta condición el término de suavidad es limitado fuertemente haciendo que la segmentación sea muy sensible al ruido. Mientras que en la energía descrita en la ecuación (6.19), la restricción de suavidad se aplica de manera homogénea a todos los sitios de la imagen, trasladando al término que maximiza la kérnel-predictibilidad la responsabilidad de hacer coincidir las interfases de la nueva segmentación con las fronteras de las regiones homogéneas de la segmentación de intensidad; más aún, en regiones finamente texturizadas el término de la kérnel-predictibilidad puede reducirse (con lo que solamente actuaría el término de suavidad), haciendo $W(|R_i|)$ directamente proporcional a la cardinalidad de R_i ($W(|R_i|) = |R_i|$ por ejemplo); con esto las regiones pequeñas tienen menos peso en la energía que las regiones con gran extensión. El maximizar la kérnel-predictibilidad dentro de una región R_i permite además acelerar la convergencia de la segmentación, pues la información de sitios en los cuales el problema de segmentación se encuentra bien definido se extiende inmediatamente al resto de los sitios de la región.

Debe notarse la generalidad de la distribución a priori definida a través de la energía gibbsiana (6.19). Aunque su aplicación puede realizarse a través de cualquier método bayesiano, específicamente utilizando el enfoque de campos aleatorios gaussianos con entropía controlada (6.15), puede obtenerse la siguiente energía:

$$\begin{aligned} U(\mathbf{p}) &= -\sum_i \left\{ \sum_k p_k^2(s_i) \left[\ln v_k(s_i) - \mu \right] \right\} + \sum_c V_c(\mathbf{p}) - \\ &\quad \lambda \sum_i W\left(|R_i|\right) \left[\sum_k \left(\frac{1}{|R_i|} \sum_{s \in R_i} p_k(s) \right)^2 \right] , \end{aligned} \tag{6.22}$$

en donde:

$$\frac{1}{|R_i|} \sum_{s \in R_i} p_k(s) \approx P(F_{R_i} = k) = \frac{1}{|R_i|} \sum_{s \in R_i} \delta(F_s - k) \; .$$

6.5. Segmentación de Disparidades.

La energía descrita en (6.21) puede aplicarse en la segmentación de disparidades existentes entre un par de imágenes estereoscópicas. Dadas dos imágenes I_L e I_R, se busca asignar una etiqueta, $l_i \in L = \{1, 2, \ldots, N\}$, a cada pixel de la imagen I_L, de manera que se cumpla la siguiente igualdad:

$$I_R(s + d_i) = I_L(s) \; , \forall s \; ,$$

en donde d_i es el valor de la disparidad definida por el modelo i-ésimo.

Suponiendo ruido gaussiano, el negativo del logaritmo de la verosimilitud para cada modelo es:

$$-\ln v_k(s) = [I_R(s + d_k) - I_L(s)]^2 \ ,$$

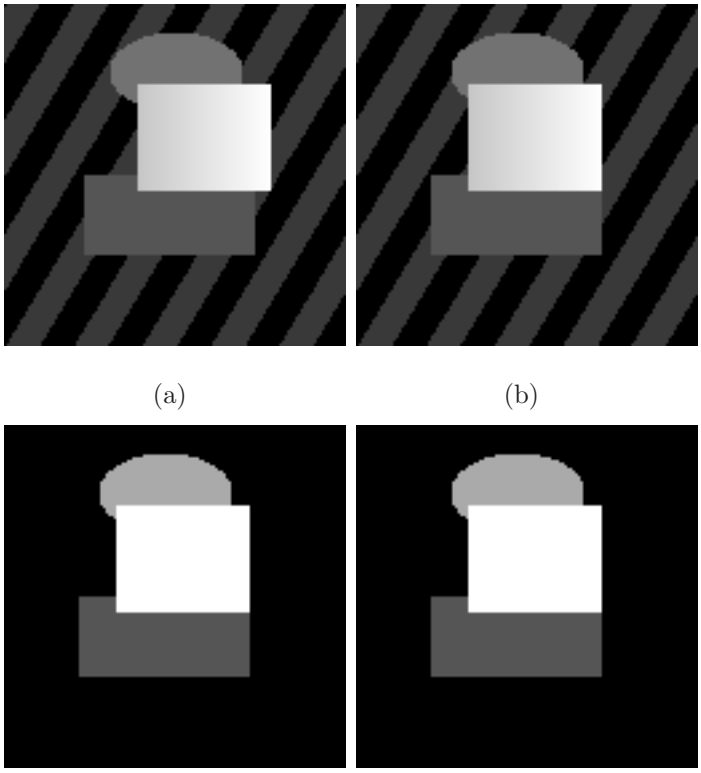
esta elección, sin embargo, vuelve la segmentación muy sensible a valores atípicos. Por lo que se justifica utilizar la siguiente expresión más robusta:

$$-\ln v_k(s) = |I_R(s + d_k) - I_L(s)| \ .$$

Sustituyendo directamente en la energía gibbsiana (6.21), la segmentación puede realizarse encontrando los valores $p_k(s), \forall(k, s)$, que la minimicen. Este proceso puede realizarse a través de descenso de gradiente.

Algunos resultados obtenidos con la aplicación de esta idea se muestran en las imágenes 6.1 y 6.2. Aquí, las regiones homogéneas que sirvieron de base para la segmentación de disparidades se obtuvieron extrayendo las regiones conexas, de una segmentación de tonos de gris con modelos constantes (35 modelos), realizada a través del método de campos de medida gaussiano-markovianos con entropía controlada. Estas regiones conexas se obtuvieron sembrando semillas sobre los puntos de la imagen segmentada y creciendo la región agregando pixeles vecinos siempre y cuando tuvieran el mismo tono de gris. Estos resultados deben contrastarse con los mostrados en la figura 6.3, en donde se ha utilizado la energía (6.22) excluyendo el término que penaliza la entropía de Gini sobre regiones homogéneas de la segmentación en tonos de gris (haciendo $\lambda = 0$). Esta comparación muestra la viabilidad de la energía propuesta, sin embargo algunos puntos deben remarcarse. Por construcción, los resultados obtenidos al utilizar la entropía de Gini regional como conocimiento a priori dependen totalmente de la segmentación de

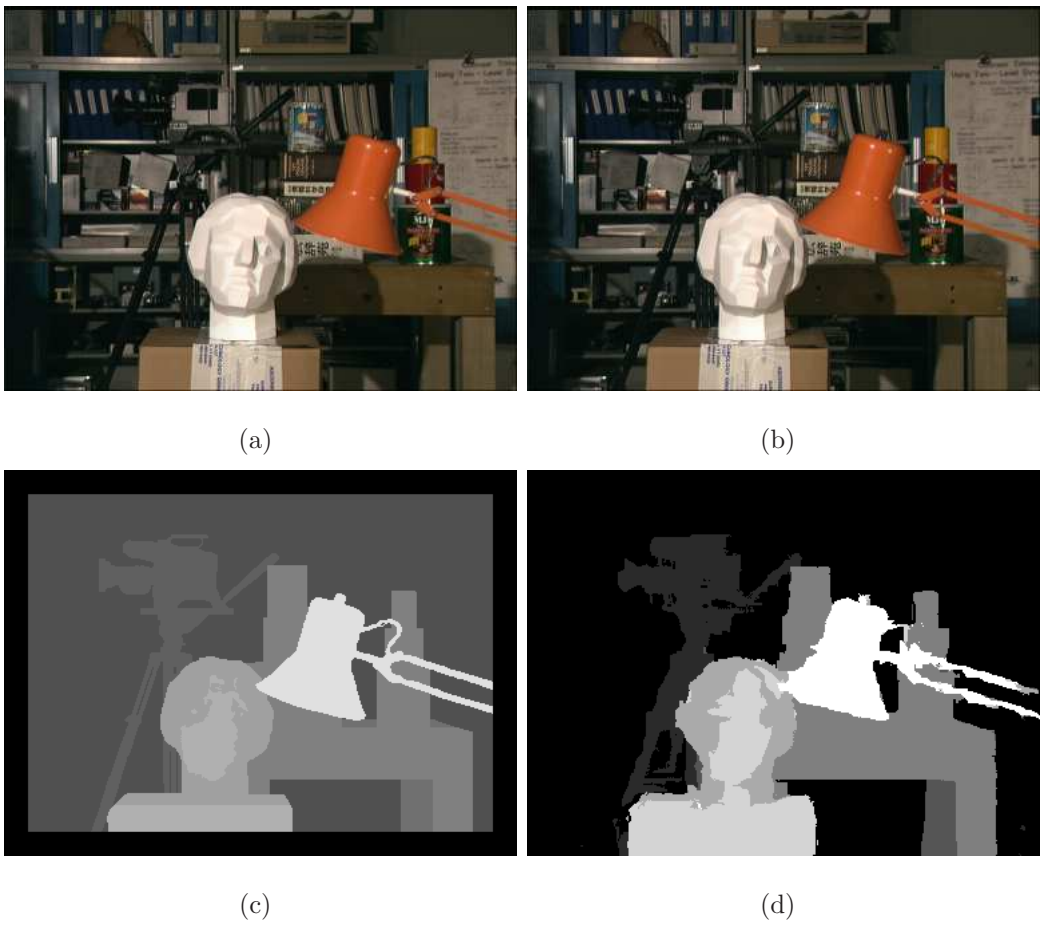
intensidades que se utilice como base. Una sub-segmentación de los tonos de gris de la imagen puede generar una segmentación de disparidad con grandes errores, al confundir las fronteras entre objetos; en el otro extremo, una sobre-segmentación de tonos de gris lleva también a definir pobremente las fronteras entre objetos, con lo que las ventajas de utilizar la entropía de Gini regional se reducen (en el extremo en el que cada pixel represente una región homogénea el término de la entropía de Gini regional no debería provocar diferencia alguna). En los ejemplos mostrados los modelos de disparidad fueron definidos mediante traslaciones constantes. Estos modelos dejan de ser apropiados cuando los objetos presentes en la escena muestran una estructura espacial más compleja. En estos casos, para representar adecuadamente la disparidad puede ser necesario aplicar varios modelos de traslación a un mismo objeto, lo cual contradice el conocimiento a priori definido en el criterio de la minimización de la entropía de Gini regional. Una alternativa podría ser el utilizar modelos que permitan un mayor número de grados de libertad, sin embargo la determinación de los parámetros que definen cada modelo es un problema cuya solución no es trivial. La solución adecuada de de estos problemas puede derivar en un método de segmentación de disparidades altamente competitivo.



(a) (b)

(c) (d)

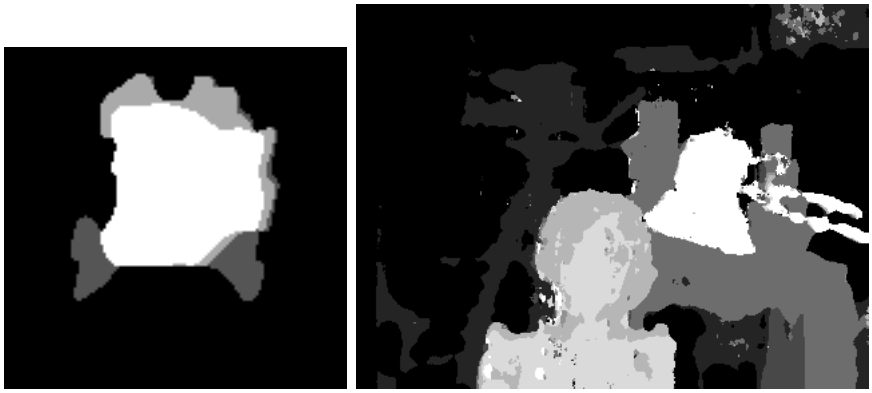
Fig. 6.1: Par estereo sintético (figs. 6.1(a) y 6.1(b)), mapa de disparidades real (fig. 6.1(c)) y mapa de disparidades obtenido (fig. 6.1(d)).



(a) (b)

(c) (d)

Fig. 6.2: Par estereo de Tsukuba (figs. 6.2(a) y 6.2(b)), mapa de disparidades real (fig. 6.2(c)) y mapa de disparidades obtenido (fig. 6.2(d)).



(a) (b)

Fig. 6.3: Resultados obtenidos sin la aplicación del término que penaliza la entropía de Gini en zonas homogéneas de la segmentación en tonos de gris.

7. CONCLUSIONES

En este trabajo se ha presentado una nueva medida de información denominada kérnel-predictibilidad, basada en el valor esperado de la evaluación de un kérnel adecuado, sobre diferentes pares de muestras de alguna distribución, generadas aleatoria e independientemente. La kérnel-predictibilidad ha sido aplicada exitosamente en el problema de registro multimodal de imágenes (tanto paramétrico como no-paramétrico), mostrando experimentalmente una gran robustez al aplicarse en problemas de registro, comparada con otras metodologías ampliamente utilizadas basadas en la entropía de Shannon como medida de información. Esta robustez se presenta principalmente en problemas de registro que envuelven transformaciones de gran magnitud y en los casos en el que el registro se realiza utilizando un conjunto reducido de muestras. Lo anterior tiene su origen en el hecho de que la kérnel-predictibilidad es una medida que depende fuertemente de los valores más significativos de la distribución de probabilidad y es poco sensible a los menos significativos, a diferencia de la entropía de Shannon. Se ha presentado también una propuesta para aplicar la kérnel-predictibilidad como conocimiento a priori en la segmentación general de imágenes, buscando minimizar el negativo de esta medida sobre regiones homogéneas en tonos de gris que pueden servir de base para una segmentación diferente. La aplicación de esta propues-

ta al problema de segmentación de la disparidad entre pares estereoscópicos se ha presentado con algunos ejemplos y es prometedora, sin embargo, aún debe extenderse la metodología para obtener un método de aplicación general y que supere el estado del arte en el área. Estas necesidades dirigen parte del trabajo a futuro. En ese sentido, otro punto a explorar, es la aplicación de la kérnel-predictibilidad en problemas de clasificación general, en los cuales la entropía de Gini ha sido utilizada anteriormente, tratando de explotar la propiedad de no-invarianza ante permutaciones de las entradas del vector de probabilidades que presenta la kérnel-predictibilidad a diferencia de la entropía de Gini.

APÉNDICE

Available online at www.sciencedirect.com
ScienceDirect
Computer Vision and Image Understanding xxx (2008) xxx–xxx
Computer Vision and Image Understanding
www.elsevier.com/locate/cviu

Image registration based on kernel-predictability ☆

Héctor Fernando Gómez-García [a,b,*], José L. Marroquín [a], Johan Van Horebeek [a]

[a] Center for Research in Mathematics (CIMAT), Computer Science, Apartado Postal 402, C.P. 36000 Guanajuato, Gto, Mexico
[b] Department of Basic Sciences and Engineering, Universidad del Caribe, C.P. 77528, Cancún Q. Roo, Mexico

Received 8 August 2006; accepted 8 February 2008

Abstract

In this work, a new similarity measure between images is presented, which is based on the concept of predictability of random variables evaluated through kernel functions. Image registration is achieved maximizing this measure, analogously to registration methods based on entropy, like mutual information and normalized mutual information. Compared experimentally with these methods in different problems, our proposal exhibits a more robust performance specially for problems involving large transformations and in cases where the registration is done using a small number of samples, such as in nonparametric registration.

Keywords: Multimodal image registration; Parametric and nonparametric transformations; Gini entropy; Information measures

1. Introduction

Due to its wide range of applications, image registration is a problem that has been largely explored (see [8,17,12] and references contained there in). It has become a fundamental task in many important fields such as robot vision and medical image processing, among others. Given a source and a reference image, represented by I_S and I_R, respectively, the registration problem consists in finding a transformation T that applied to I_S aligns it spatially to I_R. Different approaches can be followed to solve the problem; many of them are based on the assumption that the intensity of every point x in the image I_R is conserved in image I_S but at a different spatial position $T(x)$. This means that the equality $I_S[T(x)] = I_R(x)$ holds for every point in I_R (known as the *Optical Flow Constraint*), and there is a huge number of registration methods based on it [11,16,21,22,2].

Hereafter, we denote by I_T the transformed source image, that is $I_S[T(x)] = I_T(x)$.

The optical flow constraint is not always applicable, for example, when registering medical images obtained from different modalities. For this case, registration by the maximization of *Mutual Information* (*MI*) has been widely used because it does not assume a functional relationship between the intensities of the images; instead, it is based on the fact that if aligned, the maximal dependency (information) between the intensities is found.

Given two images, I_T and I_R, their mutual information is defined as:

$$MI(I_T, I_R) = H(I_T) + H(I_R) - H(I_T, I_R) \quad (1)$$

where H is the entropy function of the image intensities. If the space of intensity values is discrete, then the entropy function is written as:

$$H(I) = -\sum_i p_i \log p_i \quad (2)$$

where p_i is the probability to observe the intensity value i; and in case of a continuous space, the entropy is defined as:

$$H(I) = -\int_{-\infty}^{\infty} p(i) \log[p(i)] \mathrm{d}i. \quad (3)$$

☆ The authors were partially supported by Grant 46270 of CONACyT (Consejo Nacional de Ciencia y Tecnología, México).

* Corresponding author. Address: Center for Research in Mathematics (CIMAT), Computer Science, Apartado Postal 402, C.P. 36000 Guanajuato, Gto, Mexico.
E-mail addresses: hector@cimat.mx (H.F. Gómez-García), jlm@cimat.mx (J.L. Marroquín), horebeek@cimat.mx (J. Van Horebeek).

1077-3142/$ - see front matter © 2008 Published by Elsevier Inc.
doi:10.1016/j.cviu.2008.02.001

The first applications of *MI* to the image registration problem, were published simultaneously by Viola et al. [23] and Collignon et al. [3], both in the middle of the last decade. Since then, a great number of publications has appeared extending the initial work to problems like nonparametric multimodal image registration [10,4], registration of stereoscopic pairs [7,13] or feature tracking in images [5].

In general, methods based on the maximization of *MI*, start with an initial transformation T^0, leading to a *MI* value MI^0, and using a proper optimization method, a sequence of transformations is generated in such a way that the associated *MI* is increased until convergence. During the optimization process, the increments in *MI* are calculated with the expression:

$$\Delta MI = \Delta H(I_T) + \Delta H(I_R) - \Delta H(I_T, I_R).$$

If the discrete version of the entropy (2) is considered, this is a function of the entries of the probability vector; hence, using a Taylor series expansion, a linear approximation for the increment in entropy is given by:

$$\Delta H = -\sum_i [1 + \log p_i]\Delta p_i.$$

Because the coefficient $[1 + \log p_i]$ is large for small probability values, this increment is highly determined by small features in the images to be registered (which are generally associated with small probability values). This can trap the registration algorithm in local optima when aligning small features, particularly if the small probabilities are not accurately computed. This makes it difficult to apply *MI* in cases where only a limited sampling is available, for example when measuring entropy at a local level in images, which is important in interesting problems like nonparametric image registration, and in the segmentation of motion between frames, where local measurements must be taken in order to learn the local motion models and to have enough spatial definition at the motion interfaces.

Another problem related to the application of *MI*, occurs when working with images with a large background compared to the region of interest, as frequently happens in medical image problems. Under this circumstance the sum of the marginal entropies can become larger than the joint entropy, leading to an increase of *MI*, instead of decreasing it in misregistration. Studholme et al. [20] proposed the use of a normalized version of the *MI* to overcome this disadvantage. This measure is known as *Normalized Mutual Information* (*NMI*):

$$NMI(I_T, I_R) = \frac{H(I_T) + H(I_R)}{H(I_T, I_R)}. \quad (4)$$

In this work we propose a new criteria for the registration of images with different intensity structure (e.g., medical images in different modalities) which uses a new predictability measure for probability distributions, which we call *Kernel-Predictability* (*KP*). *KP*, evaluated in the marginal and joint distributions of two images, is integrated in a similarity measure between images, normalized as (4), and applied to the registration problem. Unlike entropy, the increment of this measure when updated by an iterative optimization method, is mostly determined by the larger entries of the probability vector, which is reflected in a higher robustness in problems where only limited sampling is available. Our proposal is discussed in Sections 2 and 3, and in Section 4 its performance in image registration problems is compared to that obtained under maximization of *MI* and *NMI*. The experimental results show that an important reduction in registration errors is obtained by the use of our method compared to *MI* and *NMI*.

2. Kernel-predictability

In order to introduce our predictability measure for a given distribution F, consider the following guessing game: someone generates a value x_1 from F and we guess x_1 by generating (independently) another value x_2 from F. We denote by $K(x_1, x_2)$ the obtained reward. Repeating this game, we can define the average reward $E[K(X_1, X_2)]$. We suppose that the reward function favors guesses close to the true value, i.e., K is a decreasing function of the distance between x_1 and x_2. Under this assumption it is clear that the less uncertainty is contained in F, the higher will be the average reward.

The above motivates the following measure for a given distribution F:

$$KP(F) = E[K(X_1, X_2)] = \int_{R^d}\int_{R^d} K(x_1, x_2)\mathrm{d}F(x_1)\mathrm{d}F(x_2). \quad (5)$$

This functional measures the predictability of the random variables distributed according to F, weighted by the kernel function K, and we denominate it *kernel-predictability*. It should be noted that *KP* is a predictability measure, so it behaves in an inverse way compared to entropy, which is an uncertainty measure.

For the discrete case, this becomes:

$$KP(\mathbf{p}) = \sum_i \sum_j K_{ij} p_i p_j = \mathbf{p}^{\mathrm{T}}\mathbf{K}\mathbf{p} \quad (6)$$

where the entry (i, j) of the matrix $\mathbf{K}$ equals the reward given for guessing the value x_i when the generated value was x_j, i.e., $K_{ij} = K(x_i, x_j)$. In the past, some measures have been presented that are apparently similar to our proposal. However an important difference must be noted. In [25], a functional like (5) is used to compute the expected distance between two groups of images. In [24,19], similarity measures between images are presented that can be confused with one of the estimators for (5) (discussed below). However, these three measures are evaluated over two different distributions, in contrast to (5), which takes only one distribution for its argument and therefore represents a property of the underlying distribution, such as its entropy or its variance.

Please cite this article in press as: H.F. Gómez-García et al., Image registration based on kernel-predictability, Comput. Vis. Image Understand. (2008), doi:10.1016/j.cviu.2008.02.001

We can measure the increment in kernel-predictability, which may be associated to the optimization process as:

$$\Delta KP = 2\sum_i \left(\sum_j K_{ij}p_j\right)\Delta p_i.$$

Note that the increment for every element of the probability vector, p_i, is multiplied by the coefficient $(\sum_j K_{ij}p_j)$; this coefficient equals the ith element of the vector generated by the product of the matrix $\mathbf{K}$ with the distribution vector $\mathbf{p}$. This product just smooths the probability vector $\mathbf{p}$ if we assume that the closer K_{ij} is to the main diagonal, the higher its value. Consequently, $(\sum_j K_{ij}p_j)$ is larger for large p_i values, and the increment in KP is mainly determined by the larger entries of the probability vector, and for that reason, by the most important features in the images to be registered. This is an important difference with respect to entropy.

2.1. Kernel-predictability with Gaussian kernels

Many choices for K are possible; a natural one is the Gaussian kernel, which is defined as:

$$K(x_1, x_2) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right) \quad (7)$$

where d is the dimension of the distribution and σ a free parameter.

For an arbitrary continuous distribution F, one can build a nonparametric approximation of its density by means of gaussian windows [6], centered over a set of points $\{a_i\}$ (e.g., independent samples obtained from F):

$$f_X(x) = \frac{1}{N}\sum_{i=1}^{N} f_{a_i,\sigma_2}(x) \quad (8)$$

where $f_{a_i,\sigma_2}(x)$ is the multivariate gaussian density, $\mathcal{N}(a_i, \sigma_2^2\mathbf{I})$, with a $d \times d$ covariance matrix $\sigma_2^2\mathbf{I}$. Moreover, if one uses a multivariate gaussian kernel to measure $KP(F)$, using the fact that a convolution of two Gaussians is another Gaussian, one can show that:

$$KP(F) = \frac{1}{N(2\pi(\sigma^2 + 2\sigma_2^2)^{d/2}} \times \sum_i \sum_j \exp(-||a_i - a_j||^2/2(\sigma^2 + 2\sigma_2^2)). \quad (9)$$

Note that the higher the spread of the points $\{a_i\}$ in the distribution, the lower will be its KP value. The maximum is reached when all the points in the set $\{a_i\}$ are equal, which represents a single Gaussian distribution. In this case, the value of KP is inversely proportional to the variance σ_2 of this distribution, which implies that the maximum value of KP will be reached when σ_2 is equal to zero, i.e., if one has a degenerate random variable that can only take one fixed value. Note that this will be true for the discrete case and for an arbitrary kernel as well, provided that the elements on the main diagonal of the matrix $\mathbf{K}$ contain the maximal reward value, say K_M (given for an exact prediction). This follows from the next inequality:

$$KP(\mathbf{p}) = \sum_i \sum_j K_{ij}p_i p_j \leqslant K_M \sum_i \sum_j p_i p_j = K_M$$

and from the fact that K_M is the value obtained for such degenerate random variables (see Fig. 1).

One important difference of KP with respect to entropy also follows from Eq. (9) and is illustrated in Fig. 2(a). If we move the Gaussian window centered over a_1 towards a_1^*, i.e., if we move a portion of the mass of the distribution to a position where there is practically no overlap with the original distribution, KP will be reduced, since the spread of the set $\{a_i\}$ will increase, and the entropy will increase. However, if one moves a_1 to a point a_1^{**} which is farther to the right, KP will be reduced even more, but the entropy will remain practically constant. This property of the entropy is not an advantage when applied in problems like image registration where the quality of a spatial transformation is measured by the narrowness of the joint distribution of gray tones between a pair of images; in this case, the gradient of KP will contain more information about the location of the optimal transformation.

Constructing the matrix $\mathbf{K}$ in (6) according to the Gaussian kernel (ignoring the normalizing constant for simplic-

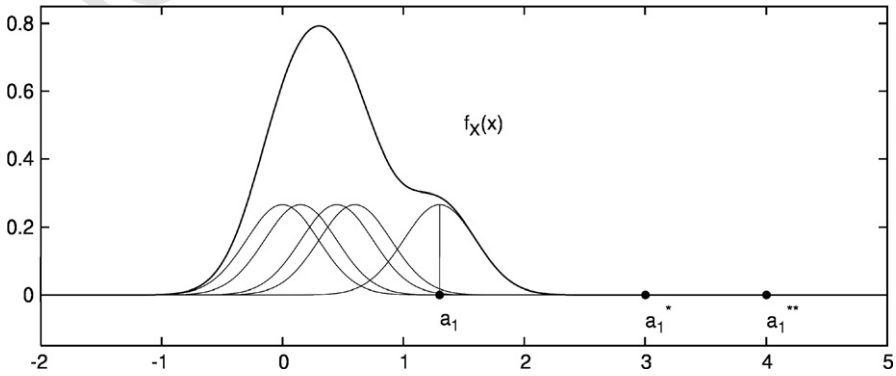


Fig. 1. Moving the gaussian window centered over a_1 towards a_1^* will reduce entropy and KP. Moving a_1 further to the right will reduce even more KP, while entropy remains constant.

Please cite this article in press as: H.F. Gómez-García et al., Image registration based on kernel-predictability, Comput. Vis. Image Understand. (2008), doi:10.1016/j.cviu.2008.02.001

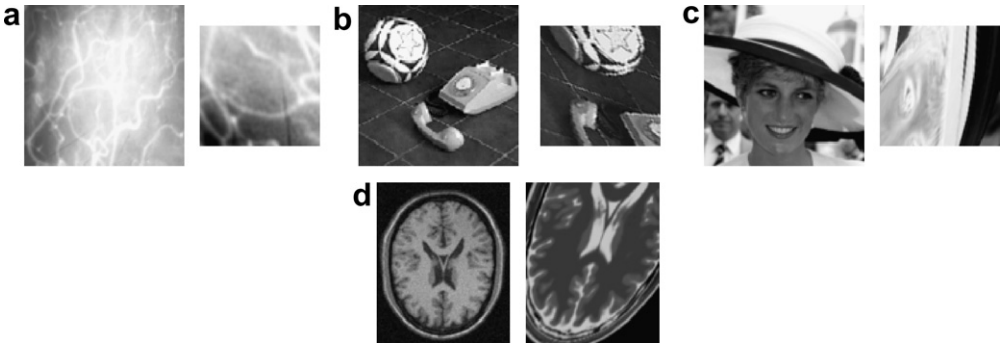


Fig. 2. Images used for registration. For cases (a)–(c), reference images were obtained applying changes in intensity and affine transformations to the original images, and then extracting a subsquare of the center of the transformed images. For case (d), the reference images were generated applying only affine transformations to a MR in modality *T*2.

ity), generates two interesting cases when evaluating the kernel at extreme values of the amplitude parameter σ. In the first case the Gaussian kernel can be approximated by the Kronecker delta for very small values of σ in the following way:

$$G(x_i, x_j) = \delta_{ij} = \begin{cases} 1 & \text{if} x_i = x_j \\ 0 & \text{otherwise.} \end{cases} \quad (10)$$

for this case, $KP(\mathbf{p}) = 1 - Gini(\mathbf{p})$, where *Gini* is the well known *Gini* entropy of Machine Learning [9]. The Gini entropy is maximized, and the associated *KP* minimized, under the uniform distribution.

For large values of σ the Gaussian kernel can be approximated by:

$$G_\sigma(x_1, x_2) \approx 1 - \frac{\|x_1 - x_2\|^2}{2\sigma^2} \quad (11)$$

and for this case, $KP(\mathbf{p}) \approx 1 - \frac{\sum_i Var[(X)_i]}{\sigma^2}$, where $Var[(X)_i]$ is the variance of the *i*th element of the multivariate random variable *X*. It can be shown that for univariate distributions with finite domain over the interval $[a, b]$, the distribution with maximal variance, and hence minimal associated *KP*, has a density equally concentrated on its two extreme values, *a* and *b*.

Random variables with uniform distribution are more difficult to predict than variables that take only two different values with the same probability, thus we prefer *KP* to behave in a way similar to the Gini entropy; for this reason we choose small values for the width of the Gaussian kernel; in practice for univariate random variables we take σ around 2–10% of their range.

2.2. Estimation of the kernel-predictability

The expression (5) is a *regular statistical functional* of degree two (two refers to the number of arguments of *K*), and for its estimation three different approaches are available in the literature [14,15]. The estimators are always based on a sampling set composed by *n* independent and identically distributed random variables, $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$, with $X_i \sim F, \forall i$; and are defined as:

$$\widehat{KP}^1 = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} K(X_i, X_j) \quad (12)$$

$$\widehat{KP}^2 = \frac{4}{n^2} \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} K(X_i, X_j) \quad (13)$$

$$\widehat{KP}^3 = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} K(X_i, X_j). \quad (14)$$

For the estimator $\widehat{KP}^1$, the kernel is evaluated over all different pairs of variables in $\mathbf{X}$; in $\widehat{KP}^2$, the set $\mathbf{X}$ is divided in two subsets and the kernel is evaluated at each pair formed by taking one variable from the first set and other variable from the second one; finally, in the third estimator, $\widehat{KP}^3$, the kernel is evaluated in all possible pairs of variables, as with estimator $\widehat{KP}^1$, but it adds the evaluations where the first and second variable coincide. The first two estimators are unbiased. If the kernel *K* is symmetric then $\widehat{KP}^1$ has the minimal variance among all the unbiased estimators, as shown in [14,15]; $\widehat{KP}^2$ has more variance than $\widehat{KP}^1$ but has a lower computational cost; the estimator $\widehat{KP}^3$ has minimal variance among these three estimators, but is biased. When the sampling set is increased in size, the variances of these estimators tend to the same value and the bias of the estimator $\widehat{KP}^3$ tends to zero.

3. Image registration with kernel-predictability

Application of *KP* to the registration problem can be done considering the joint distribution of the intensities of the images I_R and I_T, that is, $p(I_R, I_T) = p(\mathbf{I}_J(T))$. The intuitive idea is that when $T = T^*$ (the correct aligning transformation), $p(\mathbf{I}_J(T^*))$ should be more concentrated than $p(\mathbf{I}_J(T))$ for $T \neq T^*$, and therefore, $KP[p(\mathbf{I}_J(T^*))] > KP[p(\mathbf{I}_J(T))]$ for $T \neq T^*$. For example, if there exists a deterministic tone transfer function Φ, between I_R and $I_{T^*}, p(\mathbf{I}_J(T^*))$ must be ordered along a ridge-like structure determined by Φ: in this case, the conditional density $p(I_{T^*}|I_R = i) = \delta(I_{T^*} - \Phi(i))$, and any other transformation must redistribute the conditional density at different tone values. It is not enough, however, to consider only the

KP evaluated over the joint distribution of I_R and I_T, because, for example, it can be maximized under transformations that assign all points in the image I_S to a single point in I_R. Restriction over the solution space can be considered normalizing the joint *KP*, in a way similar to what is done for mutual information [20]. We propose the next similarity measure between images based on *KP*:

$$SKP(I_T, I_R) = \frac{KP[p(\mathbf{I}_J)]}{KP[p(I_T)] + KP[p(I_R)]}. \tag{15}$$

This similarity measure makes a comparison between the predictability of the joint distribution and that of the marginal distributions for the images I_T and I_R. An upper bound for *SKP* in the discrete case is derived in Appendix A. For the particular case where the kernel K used for the evaluation of *KP* is the Kronecker delta, it is possible to show rigorously that SKP reaches its global maximum for $T = T^*$ (see Appendix A). This kernel, however, is not appropriate for practical computations, because in this case the gradient of SKP has very little information about the location of its maximum. In general, at least a local maximum of SKP is obtained for Gaussian kernels as well; to see this, note that for T different, but close to T^*, $KP[p(I_T)] + KP[p(I_R)] \approx KP[p(I_{T^*})] + KP[p(I_R)]$ and $KP[p(\mathbf{I}_J(T^*))] > KP[p(\mathbf{I}_J(T))]$, since $p(\mathbf{I}_J(T))$ is less concentrated than $p(\mathbf{I}_J(T^*))$ (see Eq. (9) and discussion above). In practice, this condition holds also for smooth kernels, for which *KP* behaves very much like the Gaussian case (see Section 2.1).

Registration of the images I_S and I_R is done by searching for the transformation T which maximizes the *SKP* value between the corresponding I_T image and I_R. The transformation can be classified as *parametric* or *nonparametric*; for each case, a different registration strategy must be followed as detailed below. Assuming it is clear from the context for which images the similarity measure is evaluated, we will write $SKP(T)$ instead of the expression $SKP(I_T, I_R)$.

3.1. Parametric registration

Suppose the transformation T is determined by a vector of m real parameters, $\mathbf{a} = (a_1, a_2, \ldots, a_m)$, and m is considerably smaller than the total number of points in the images to be registered; in this case, we write $T(x; \mathbf{a})$ instead of $T(x)$ (e.g., when registering images under affine or projective transformations). For ease of notation, the intensity values associated to an arbitrary sampled coordinate X_i, can be abbreviated with the expressions: $I_R^i = I_R(X_i)$, $I_T^i = I_S[T(X_i; \mathbf{a})]$, and $\mathbf{I}_J^i = (I_T^i, I_R^i)$. Then, an approximation to (15) using the estimator (13) can be written in the following way:

$$\widehat{SKP}[T(\mathbf{a})] = \frac{\widehat{KP}_J[T(\mathbf{a})]}{\widehat{KP}_T[T(\mathbf{a})] + \widehat{KP}_R} \tag{16}$$

with

$$\widehat{KP}_J[T(\mathbf{a})] = \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} K_{\sigma_J}(\mathbf{I}_J^i, \mathbf{I}_J^j)$$

$$\widehat{KP}_T[T(\mathbf{a})] = \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} K_{\sigma_M}(I_T^i, I_T^j)$$

$$\widehat{KP}_R = \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} K_{\sigma_M}(I_R^i, I_R^j),$$

K_{σ_J} is the kernel employed to measure the predictability of the joint distribution of I_T and I_R, and K_{σ_M} for the marginal distributions of I_T and I_R. Note that the constant coefficient in the estimators can be ignored due to normalization.

For example, if Gaussian kernels are used (ignoring the normalizing constants), then:

$$K_{\sigma_J}(\mathbf{I}_J^i, \mathbf{I}_J^j) = G_{\sigma_J}(\mathbf{I}_J^i, \mathbf{I}_J^j) = \exp\left\{-\frac{\|\mathbf{I}_J^i - \mathbf{I}_J^j\|^2}{2\sigma_J^2}\right\} \tag{17}$$

$$K_{\sigma_M}(I^i, I^j) = G_{\sigma_M}(I^i, I^j) = \exp\left\{-\frac{(I^i - I^j)^2}{2\sigma_M^2}\right\}. \tag{18}$$

The maximization can be done using stochastic gradient ascent, starting with an initial transformation defined by the vector $\mathbf{a}^0$ and actualizing it with the relation:

$$\mathbf{a}^{t+1} = \mathbf{a}^t + \lambda \nabla_{\mathbf{a}} \widehat{SKP}[T(\mathbf{a}^t)]$$

with:

$$\nabla_{\mathbf{a}} \widehat{SKP}[T(\mathbf{a}^t)] = \frac{1}{\widehat{KP}_T[T(\mathbf{a}^t)] + \widehat{KP}_R} \nabla_{\mathbf{a}} \widehat{KP}_J[T(\mathbf{a}^t)] - \frac{\widehat{KP}_J[T(\mathbf{a}^t)]}{(\widehat{KP}_T[T(\mathbf{a}^t)] + \widehat{KP}_R)^2} \nabla_{\mathbf{a}} \widehat{KP}_T[T(\mathbf{a}^t)] \tag{19}$$

and in particular, when using the kernels (17) and (18), these gradients become:

$$\nabla_{\mathbf{a}} \widehat{KP}_J[T(\mathbf{a}^t)] = -\frac{1}{\sigma_J^2} \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} G_{\sigma_J}(\mathbf{I}_J^i, \mathbf{I}_J^j)(I_T^i - I_T^j)(\nabla_{\mathbf{a}} I_T^i - \nabla_{\mathbf{a}} I_T^j)$$

$$\nabla_{\mathbf{a}} \widehat{KP}_T[T(\mathbf{a}^t)] = -\frac{1}{\sigma_M^2} \sum_{i=1}^{n/2} \sum_{j=n/2+1}^{n} G_{\sigma_M}(I_T^i, I_T^j)(I_T^i - I_T^j)(\nabla_{\mathbf{a}} I_T^i - \nabla_{\mathbf{a}} I_T^j).$$

The gradient can be estimated using a different sampling set for every iteration, giving a stochastic behavior to the gradient ascent (as is proposed in [23]) which allows the optimization procedure to escape from local optima; in this sense the use of the estimator (13) is more suitable due to the fact that its higher variance introduces an additional stochastic component. Besides it has the lowest computational cost among the three options.

When working with large transformations, the part of the image I_R in the overlapping region between the two images can vary with T, and the gradient of the similarity must consider this variation. Unfortunately there is no explicit dependence of I_R on the transformation; therefore

one must approximate the gradient of the similarity by finite differences. The partial derivative of (16) with respect to any parameter a_i can be evaluated with centered finite differences as:

$$\frac{\partial \widehat{SKP}}{\partial a_i}[T(\mathbf{a}^t)] \approx \frac{\widehat{SKP}[T(\mathbf{a}^t + \epsilon_i \mathbf{e}_i)] - \widehat{SKP}[T(\mathbf{a}^t - \epsilon_i \mathbf{e}_i)]}{2\epsilon_i}, \tag{20}$$

where $\mathbf{e}_i$ is a vector with a one in the ith component and zeros in the rest, and ϵ_i is a small real value. Using this approximation, the similarity must be evaluated twice for each parameter in the transformation and because every evaluation determines a different overlapping region between the images, in order to calculate accurately the gradient, the samples used for estimation must lie in the intersection of all overlapping regions.

The use of (20) for the gradient approximation is advantageous in the case of registrations with large transformations, where the variation of I_R during the process is not negligible; otherwise one can ignore this variation and employ the simpler approach defined in (19). In this paper, the approximation (20) was employed for registration.

3.2. Nonparametric registration

To obtain a nonparametric (dense) field, the registration must find a different translation vector for each point in the images; in this case, the transformation for every pixel is defined in the following way: $T(x_i) = x_i + u_i$, $i \in \{1, \ldots, N\}$. A large amount of sampling is necessary in order to estimate accurately the complete transformation field, $\mathbf{u} = \{u_1, \ldots, u_N\}$, and the registration by the maximization of our similarity measure can be prohibitive due to its quadratic cost over the sampling size. Instead of maximizing it globally, one can restrict its evaluation to a local level, focusing on a small region around each point in the images; then we can maximize the sum of the local similarities for every point x. For example, if we consider a small squared region defined by the window W_x centered on the point x, then the local similarity will be a function only of the translation vectors associated to the points enclosed by W_x, that is the set $\mathbf{v}_x = \{u_i | i \in W_x\}$. Besides the reduction of the computational cost, evaluating the similarity at a local level can help to avoid the irregularities of the probability distributions of the intensities, which results from large spatial inhomogeneities in the intensity of the images. Also regularization of the field $\mathbf{u}$ must be considered. Therefore, for nonparametric registration, the minimization of the following energy is proposed, which is a combination of a data fidelity term, E_D, and a smoothness term, E_S:

$$E(\mathbf{u}) = E_D(\mathbf{u}) + \lambda E_S(\mathbf{u})$$

where

$$E_D(\mathbf{u}) = \sum_x \{-\widehat{SKP}_{W_x}(\mathbf{v}_x)\} \tag{21}$$

$$E_S(\mathbf{u}) = \sum_x \left\{ \sum_{x' \in N_x} \|u_x - u_{x'}\|^2 \right\} \tag{22}$$

λ is a constant which controls the smoothness of the field, and N_x is a small neighborhood around the point x.

The local similarity is evaluated in the following way:

$$\widehat{SKP}_{W_x}(\mathbf{v}_x) = \frac{\widehat{KP_J}(\mathbf{v}_x)}{\widehat{KP_T}(\mathbf{v}_x) + \widehat{KP_R}(x)} = \frac{\sum_{i,j \in W_x} K_{\sigma_J}(\mathbf{I}_J^i, \mathbf{I}_J^j)}{\sum_{i,j \in W_x} K_{\sigma_M}(I_T^i, I_T^j) + \sum_{i,j \in W_x} K_{\sigma_M}(I_R^i, I_R^j)}. \tag{23}$$

For this case $I_T^i = I_S(x_i + u_i)$. We have written the $\widehat{KP}_R$ value as a function of the centering point, x, in order to stress its local evaluation. Note that now the estimator (14) is being used; this is due to the fact that when working with small windows, only a few samples are available for the estimation of the similarities, and the smaller variance of (14) allows for a more accurate calculation of the field; the estimator (12) can be used as well with little difference in the results, but the use of estimator (13) should be avoided, mostly for very small windows (e.g., windows with 3×3 pixels).

The minimization is done by gradient descent. When using the Gaussian kernels (17) and (18), the partial derivative of the data fidelity term in Eq. (21) with respect to any translation vector u_l is:

$$\frac{\partial E_D}{\partial u_l} = 2 \sum_{x: l \in W_x} \sum_{i \in W_x} \left\{ \begin{matrix} f_J(x) G_{\sigma_J}(\mathbf{I}_J^l, \mathbf{I}_J^i) - \\ f_M(x) G_{\sigma_M}(I_T^l, I_T^i) \end{matrix} \right\} (I_T^l - I_T^i) \nabla I_S(x_l + u_l) \tag{24}$$

where: $f_J(x) = \frac{1}{\sigma_J^2 [\widehat{KP_T}(\mathbf{v}_x) + \widehat{KP_R}(x)]}$, $f_M(x) = \frac{\widehat{KP_J}(\mathbf{v}_x)}{\sigma_M^2 [\widehat{KP_T}(\mathbf{v}_x) + \widehat{KP_R}(x)]^2}$, and $\nabla I_S(x_l + u_l)$ is the spatial gradient of the image I_S evaluated at the point $(x_l + u_l)$. Note that the first sum runs over every window, W_x, containing the point l, and the second one runs over every point within the window W_x.

Finally, the gradient of the smoothness term is:

$$\frac{\partial E_S}{\partial u_l} = 4 \left(|N_l| u_l - \sum_{l' \in N_l} u_{l'} \right).$$

Image registration by the use of (24) can be time consuming for large windows (e.g., 7×7 pixels or more). Supposing that a local kernel-predictability has been evaluated for a given point x and for a fixed set of vectors $\mathbf{v}_x^0$, then it is possible to make an approximation to evaluate the kernel-predictability for a new set of vectors $\mathbf{v}_x$, making a linear approximation in Taylor series around $\mathbf{v}_x^0$ in the following way:

$$\widehat{KP}(\mathbf{v}_x) \approx \widehat{KP}(\mathbf{v}_x^0) + (\mathbf{v}_x - \mathbf{v}_x^0)^{\mathrm{T}} \nabla_{\mathbf{v}} \widehat{KP}(\mathbf{v}_x^0). \tag{25}$$

Once the values of $\widehat{KP}(\mathbf{v}_x^0)$ and $\nabla_{\mathbf{v}}\widehat{KP}(\mathbf{v}_x^0)$ are evaluated, the approximation to the kernel-predictability is reduced from $|W|^2$ kernel evaluations, to the calculation of a product of two vectors containing $|W|$ elements without any kernel evaluation. Substituting the linearized approximations for $\widehat{KP}_J(\mathbf{v}_x)$ and $\widehat{KP}_T(\mathbf{v}_x)$ in (23), it can be rewritten as:

$$\widehat{SKP}_{W_x}(\mathbf{v}_x) = \frac{\widehat{KP}_J(\mathbf{v}_x^0) + (\mathbf{v}_x - \mathbf{v}_\mathbf{x}^\mathbf{0})^{\mathrm{T}}\nabla_{\mathbf{v}}\widehat{KP}_J(\mathbf{v}_x^0)}{\widehat{KP}_T(\mathbf{v}_x^0) + (\mathbf{v}_x - \mathbf{v}_x^0)^{\mathrm{T}}\nabla_{\mathbf{v}}\widehat{KP}_T(\mathbf{v}_x^0) + \widehat{KP_R}(x)}. \tag{26}$$

Substitution of (26) into the term (21) simplifies the gradient of the data fidelity term to:

$$\frac{\partial E_D}{\partial u_l} = -\sum_{x:l\in W_x}\{f_J(x)[\nabla_{\mathbf{v}}\widehat{KP}_J(\mathbf{v}_x^0)]_l - f_M(x)[\nabla_{\mathbf{v}}\widehat{KP}_T(\mathbf{v}_x^0)]_l\} \tag{27}$$

where $[\nabla_{\mathbf{v}}\widehat{KP}_J(\mathbf{v}_x^0)]_l$ and $[\nabla_{\mathbf{v}}\widehat{KP}_T(\mathbf{v}_x^0)]_l$, are the lth component of the kernel-predictability gradients:

$$[\nabla_{\mathbf{v}}\widehat{KP}_M(\mathbf{v}_x^0)]_l = -\frac{2}{\sigma_M^2}\sum_{i\in W_x} G_{\sigma_M}(I_T^l, I_T^i)(I_T^l - I_T^i)\nabla I_S[x_l + (\mathbf{v}_x^0)_l]$$

$$[\nabla_{\mathbf{v}}\widehat{KP}_J(\mathbf{v}_x^0)]_l = -\frac{2}{\sigma_J^2}\sum_{i\in W_x} G_{\sigma_J}(\mathbf{I}_J^l, \mathbf{I}_J^i)(I_T^l - I_T^i)\nabla I_S[x_l + (\mathbf{v}_x^0)_l]$$

and $I_T^i = I_S[x_i + (\mathbf{v}_x^0)_i]$, $f_J(x) = \frac{1}{\widehat{KP}_T(\mathbf{v}_x^0) + (\mathbf{v}_x - \mathbf{v}_x^0)^{\mathrm{T}}\nabla_{\mathbf{v}}\widehat{KP}_T(\mathbf{v}_x^0) + \widehat{KP_R}(x)}$, $f_M(x) = \frac{\widehat{KP}_J(\mathbf{v}_x^0) + (\mathbf{v}_x - \mathbf{v}_x^0)^{\mathrm{T}}\nabla_{\mathbf{v}}\widehat{KP}_J(\mathbf{v}_x^0)}{[\widehat{KP}_T(\mathbf{v}_x^0) + (\mathbf{v}_x - \mathbf{v}_x^0)^{\mathrm{T}}\nabla_{\mathbf{v}}\widehat{KP}_T(\mathbf{v}_x^0) + \widehat{KP_R}(x)]^2}$.

The optimization by gradient descent using (27), requires a periodical reevaluation of the values and gradients of the kernel-predictability, in practice, after every 5–10 iterations. Using this approach an important reduction in the convergence time is reached without loosing too much accuracy.

4. Results

In this section we present some results obtained with the application of our proposal to different image registration problems.

4.1. Parametric registration

In the first set of experiments we compared the performance of our method with respect to registration by maximization of mutual information and normalized mutual information, in affine registration problems. For these measures, two different implementations were considered. The first one, uses the discrete version of the entropy (2), approximating the probability distributions by normalized histograms, and performing the optimization with the simplex method [18]; this implementation is widely used and its advantages over other implementations (in all cases using the discrete version of the entropy) are documented by Zhu and Cochoff [26]. The second implementation is based on the continuous version of the entropy (3), using Parzen windows for the estimation of the probability densities, and following [23] for the entropy estimation; these approximations are:

$$H(I_R) = -\frac{1}{|A|}\sum_{i\in A}\log\left\{\frac{1}{|B|}\sum_{j\in B}G_{\sigma_M}(I_R^i - I_R^j)\right\} \tag{28}$$

$$H[I_L(T)] = -\frac{1}{|A|}\sum_{i\in A}\log\left\{\frac{1}{|B|}\sum_{j\in B}G_{\sigma_M}(I_T^i - I_T^j)\right\} \tag{29}$$

$$H[I_L(T), I_R] = -\frac{1}{|A|}\sum_{i\in A}\log\left\{\frac{1}{|B|}\sum_{j\in B}G_{\sigma_J}(I_J^i - I_J^j)\right\}, \tag{30}$$

where A and B, are two different sets of sampled coordinates in the overlapping region of the images, and G_σ, is the normal density with variance σ^2; the optimization is done using stochastic gradient ascent, approximating the partial derivatives with centered finite differences.

Affine transformations can be applied multiplying a squared matrix $\mathbf{A}$ with a point $\mathbf{p}$ and adding a translation vector $\mathbf{t}$, to generate a transformed point $\mathbf{p}\prime$. The matrix $\mathbf{A}$ is a composition of three simpler transformations: a rotation $\mathbf{R}$, a scaling $\mathbf{S}$, and a shearing $\mathbf{H}$; this is represented by:

$$\mathbf{p}' = \mathbf{A}\mathbf{p} + \mathbf{t} = (\mathbf{RSH})\mathbf{p} + \mathbf{t}.$$

The order of the matrices multiplication is arbitrary, and for bidimensional transformations the exact representation for each matrix is:

$$\mathbf{R} = \begin{pmatrix}\cos\phi & -\sin\phi\\ \sin\phi & \cos\phi\end{pmatrix}$$

$$\mathbf{S} = \begin{pmatrix}\alpha & 0\\ 0 & \beta\end{pmatrix}$$

$$\mathbf{H} = \begin{pmatrix}1 & \gamma\\ 0 & 1\end{pmatrix}\begin{pmatrix}1 & 0\\ \delta & 1\end{pmatrix}.$$

Five sets, composed of 50 affine transformations each one, were generated assigning random values for the $\phi, \alpha, \beta, \gamma$, and δ parameters, and for the translation vector. These values were sampled uniformly from certain intervals, as is summarized in Table 1.

Different bidimensional images were used for registration (see Fig. 2). Reference images were created applying a change in intensity and affine transformations to the original images (128×128 pixels), and then extracting a square of 90×90 pixels from the center of the transformed

Table 1
Composition of the five transformations sets

Set	ϕ (degrees)	α, β	γ, δ	t (pixels for each component)
$S1$	$[-10°, 10°]$	$[0.9, 1.1]$	$[-0.1, 0.1]$	$[-10.0, 10.0]$
$S2$	$[-20°, 20°]$	$[0.8, 1.2]$	$[-0.2, 0.2]$	$[-20.0, 20.0]$
$S3$	$[-30°, 30°]$	$[0.7, 1.3]$	$[-0.3, 0.3]$	$[-30.0, 30.0]$
$S4$	$[-40°, 40°]$	$[0.6, 1.4]$	$[-0.4, 0.4]$	$[-40.0, 40.0]$
$S5$	$[-50°, 50°]$	$[0.5, 1.5]$	$[-0.5, 0.5]$	$[-50.0, 50.0]$

The width of the generating interval for each parameter is progressively augmented.

images, as is shown in Fig. 2(a)–(c), excepting images 2(d) (217×181 pixels), which correspond to two magnetic resonances obtained by the simulator at the Montreal Neurological Institute [1]; for this case, the floating image is a $T1$-weighted MRI with 9% of noise level and 40% of spatial inhomogeneities in intensity, and the reference images were created applying affine transformations to a corresponding $T2$-weighted image. The intensities of every image pair were scaled between 0 and 100; after that, the change in intensity was applied through the function $I_R = 100(\frac{I_L}{100})^{1.35}$ for images 2(a) and (b) and $I_R = 100(1 - \frac{I_L}{100})^{1.35}$ for 2(c). This process was repeated for every transformation in each set, and the algorithms executed for registering the original images to the reference images. For every registration, two Gaussian pyramids of three levels were constructed by alternatively smoothing (with a Gaussian kernel) and sub-sampling the original source and reference images; then, the registration started with the identity transformation in the coarsest level of the pyramids and the resulting transformation for every level was used as the initial transformation for the subsequent level. The implementation details for the two discrete algorithms were set according to Zhu and Cochoff [26]. For the case of continuous entropy, two different sets of coordinates composed of 50 samples each one were used. A multiple of the identity matrix, $\sigma^2 I$, was used as the covariance matrix in the estimation of the joint entropy of images, and for the marginal entropies the variance was set to the value σ^2; this value was fixed manually, considering a percentage of the dynamic range of the images to be registered. The values used in these experiments were $\sigma = 5\%$ for image 2(a) and $\sigma = 10\%$ for the rest of the images. In the case of *SKP*, estimator (13) was employed, using the same number of samples for estimation as was done with *MI* and *NMI*, and the width of the kernels used were set with the same considerations, except that a fixed value of $\sigma = 8\%$ was used for all registrations. The number of successful registrations for each set and for each algorithm, is plotted in Fig. 3; a registration was considered successful if the mean error between the applied and recovered vector fields was lower than one pixel. It can be noted that, almost in all cases, our method outperformed all versions of registration by mutual information and normalized mutual information, specially for large transformations; and that the algorithms based on the discrete version of the entropy have no robustness when used for registrations with large transformations.

Considering the algorithms based on the continuous estimation of entropy, our method presents another advantage. Due to the quadratic cost of the estimation of both kernel-predictability and entropy, a very important parameter is the number of samples used for registration; the

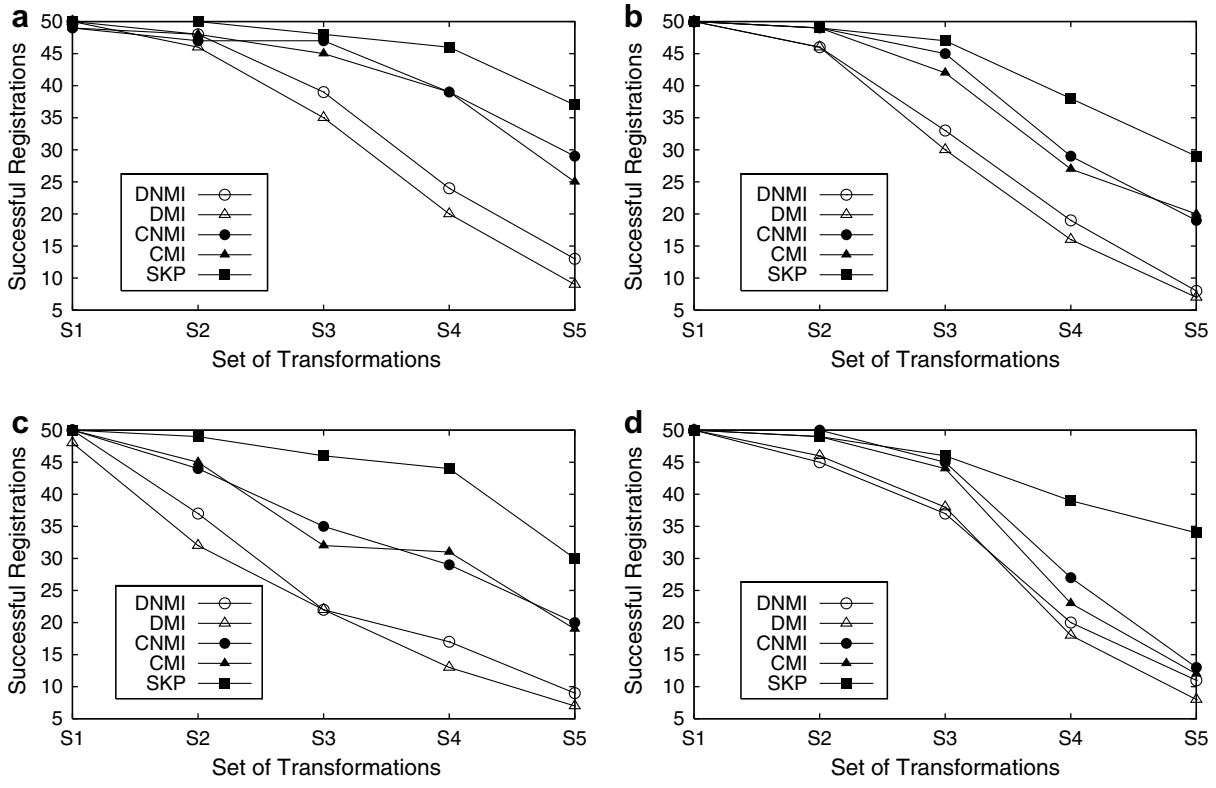


Fig. 3. Successful registrations in function of the complexity of the transformations. The plots show results corresponding to images 2(a)–(d). In the plot *SPK* means "Similarity based on Kernel-Predictability", *CNMI* and *DNMI* refers to "Continuous" and "Discrete Normalized Mutual Information"; finally *CMI* and *DMI*, refers to "Continuous" and "Discrete Mutual Information".

Please cite this article in press as: H.F. Gómez-García et al., Image registration based on kernel-predictability, Comput. Vis. Image Understand. (2008), doi:10.1016/j.cviu.2008.02.001

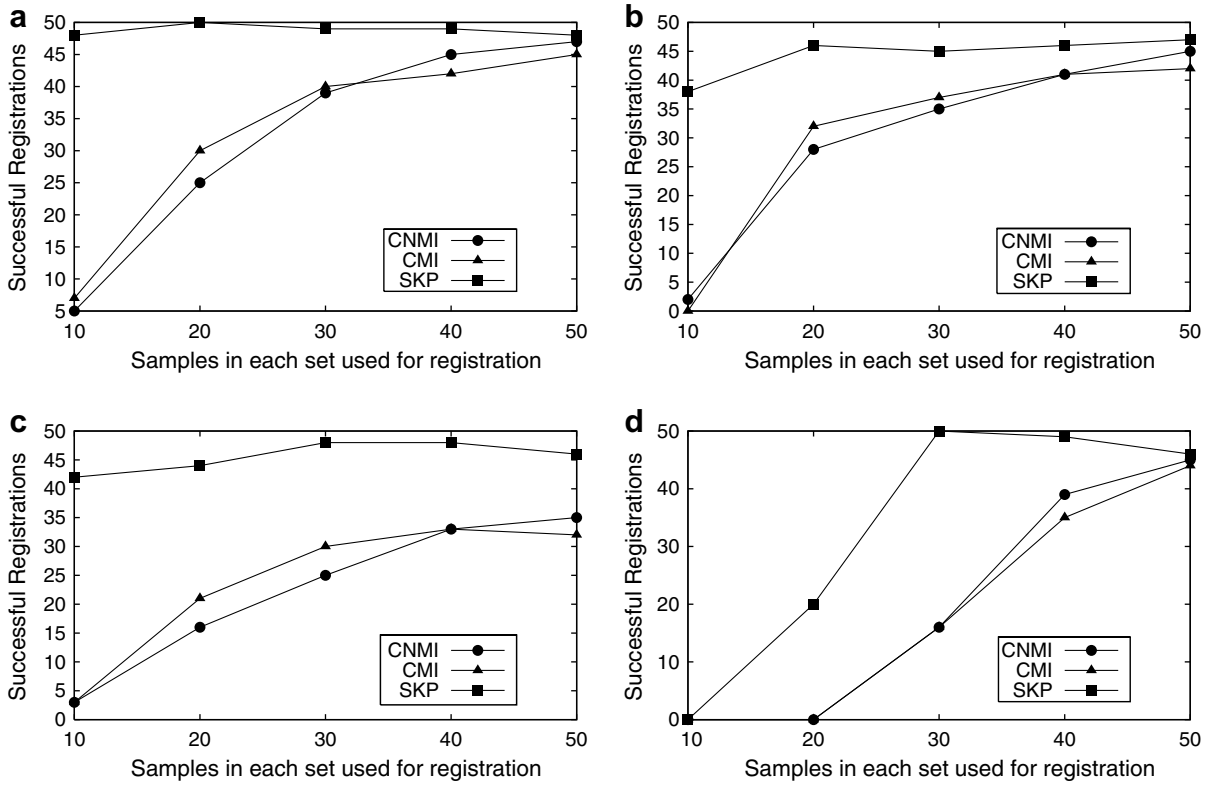


Fig. 4. Successful registrations as a function of the number of samples used for estimation. The plots show results corresponding to images 2(a)–(d). In the plot *SKP* means "Similarity based on Kernel-Predictability", *CNMI* refers to "Continuous Normalized Mutual Information"; finally *CMI* refers to "Continuous Mutual Information".

plots in Fig. 4 show the performance of the three methods when varying this parameter, in this case the set $S3$ of affine transformations (described in Table 1) was used in the four images; as can be seen, our method works considerably well even using a very small sampling for estimation, which is not the case of mutual information and normalized mutual information.

Finally, the performance of our proposal was evaluated under different kernel functions. Registration of the four image pairs (shown in Fig. 2) was repeated for *SKP* using the one-dimensional kernels described in Table 2 for the evaluation of the marginal *KP*'s. The joint *KP* was evaluated in each case, employing a separable kernel generated by the product of the two marginal kernels, that is $K_J(\mathbf{I}_J^i, \mathbf{I}_J^j) = K_M(I_R^i, I_R^j)K_M(I_T^i, I_T^j)$. It can be noted in Fig. 5(a)–(d) that the selection of the kernel for registration by maximization of *SKP* is not a critical factor. Small differences were obtained for different smooth kernels, however a poor performance is obtained in the case of the triangular kernel.

4.2. Nonparametric registration

The robustness of our proposal for working correctly with large transformations and using only few samples, makes it very suitable to be applied in nonparametric registration problems. In order to measure the performance of $\widehat{SKP}$ in these problems, 10 different synthetic transformation fields were generated using two grids with 15×15 nodes of cubic B-spline functions, and assigning random values to every node. Then, for each pixel (x, y) in the image, a translation vector $(u(x, y), v(x, y))$ was defined in the following way:

$$u(x,y) = \sum_{i=1}^{15}\sum_{j=1}^{15} U_{ij}\beta[k_1(x-x_i)]\beta[k_2(y-y_j)]$$
$$v(x,y) = \sum_{i=1}^{15}\sum_{j=1}^{15} V_{ij}\beta[k_1(x-x_i)]\beta[k_2(y-y_j)] \quad (31)$$

where $U_{ij}, V_{ij} \sim U\{-7, 7\}$, for all centering nodes (x_i, y_j), and k_d is the proportion of nodes versus the image dimen-

Table 2
Different kernels used for registration with *SKP*

Gaussian kernel	$K(x_1,x_2) = \exp[-(x_1-x_2)^2/\sigma^2]$
Cauchy kernel	$K(x_1,x_2) = \frac{1}{1+\alpha(x_1-x_2)^2}$
Exponential kernel	$K(x_1,x_2) = \exp(-\lvert x_1-x_2\rvert/\sigma^2)$
Triangular kernel	$K(x_1,x_2) = 1-\alpha\lvert x_1-x_2\rvert$ for $\alpha\lvert x_1-x_2\rvert < 1$ and $K(x_1,x_2) = 0$, otherwise

Please cite this article in press as: H.F. Gómez-García et al., Image registration based on kernel-predictability, Comput. Vis. Image Understand. (2008), doi:10.1016/j.cviu.2008.02.001

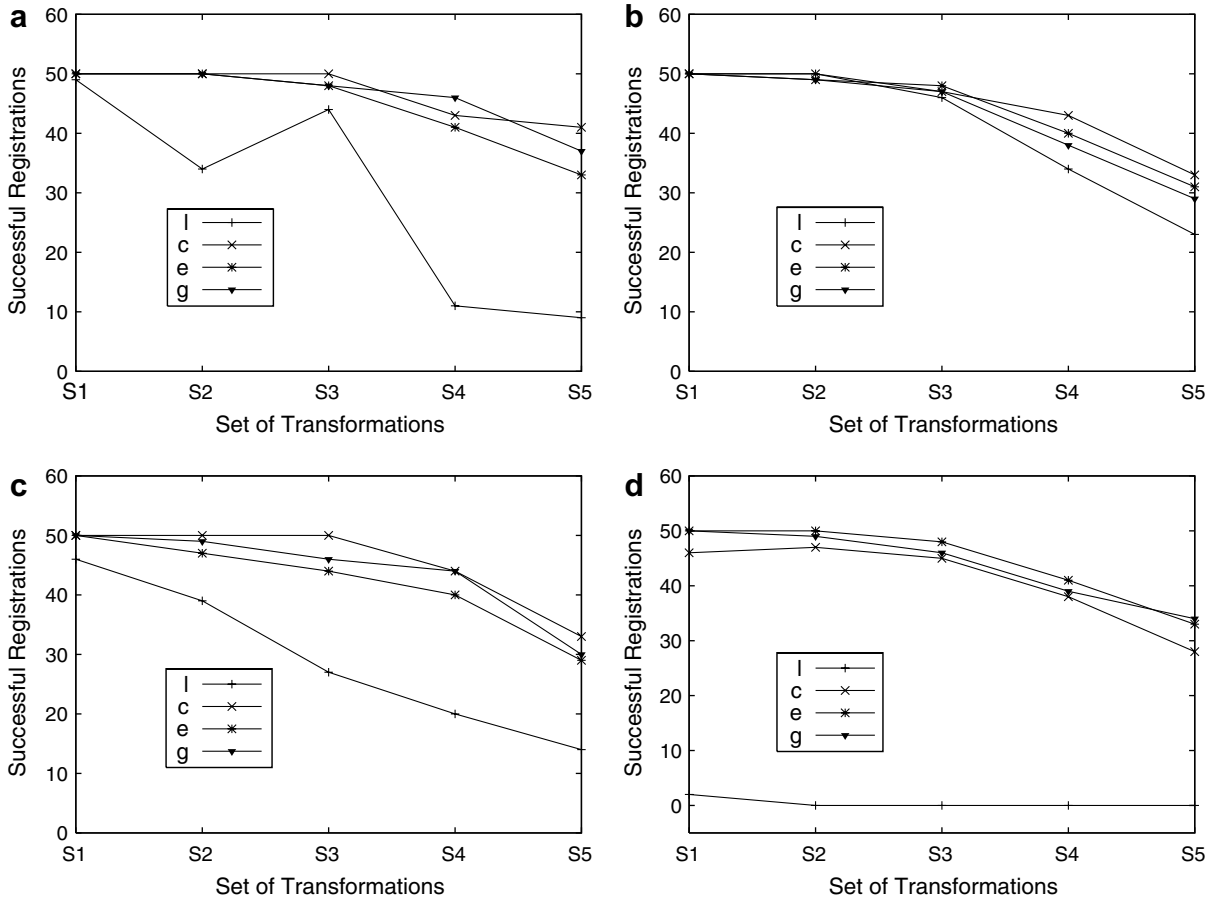


Fig. 5. Registration results for *SKP* using different kernels (described in Table 2). The plot shows registration results using *SKP* with Gaussian (g), Cauchy (c), exponential (e) and triangular kernels (l).

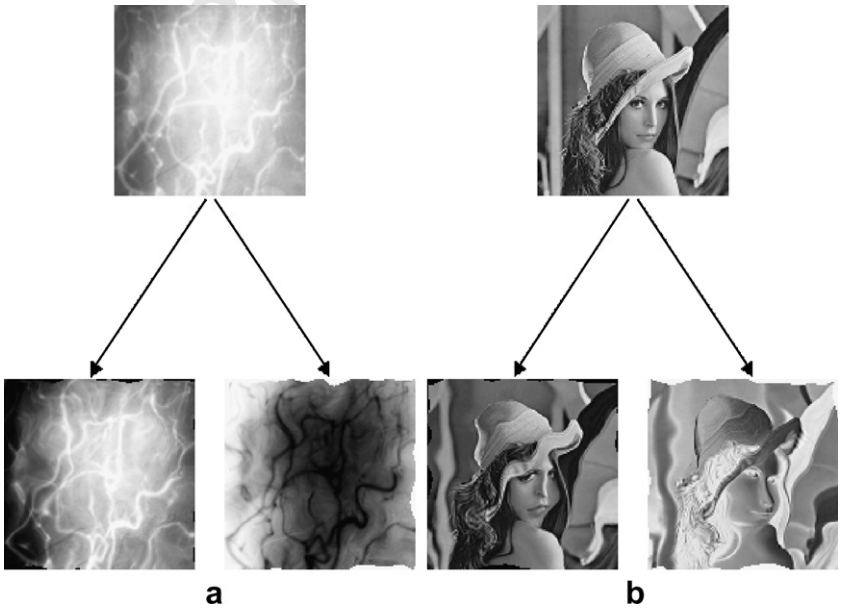


Fig. 6. Images used for nonparametric registration. Reference images were created applying changes in intensity and different synthetic transformation fields to the original images.

Please cite this article in press as: H.F. Gómez-García et al., Image registration based on kernel-predictability, Comput. Vis. Image Understand. (2008), doi:10.1016/j.cviu.2008.02.001

sion in the direction d. The cubic B-spline functions used are:

$$\beta(z) = \begin{cases} \frac{2}{3} - |z|^2 + \frac{|z|^3}{2}, & |z| < 1. \\ \frac{(2-|z|)^3}{6}, & 1 \leqslant |z| < 2 \\ 0, & |z| \geqslant 2. \end{cases}$$

The synthetic fields were applied to two images after a change in intensity determined by two different tone transfer functions, $f_1(I) = 100(\frac{I}{100})^{1.35}$ and $f_2(I) = 100(1 - \frac{I}{100})^{1.35}$ for every image, as shown in Fig. 6. Then, our nonparametric registration algorithm was executed to recover the original transformation field and the error measured for each case. The error was calculated as the average length of the difference between the applied and recovered vectors for all pixels. As was done with parametric registration, Gaussian pyramids of three levels were used for the source and reference images; in the coarsest level of the pyramids every vector of the transformation field was initialized to zero and for all the subsequent levels, the transformation was started with the resulting field of the previous level. For comparison, the registration algorithm was run substituting $\widehat{SKP}$ in the term (21) by the corresponding expressions for *MI* and *NMI* based on the continuous entropy (Eqs. (28)–(30)). As described in Section 3.2, the similarity measures were evaluated at a local level using small windows placed over each pixel in the images, and windows of different sizes were considered. The results are summarized in Fig. 7(a)–(d); as can be seen, important reductions in the mean error are obtained with our proposal compared to *MI* and *NMI* when using small windows for registration, and again, due to the quadratic cost of the estimations over the number of samples, this is reflected in important savings in the execution time (see Fig. 8). To facilitate a qualitative comparison of the errors, the regis-

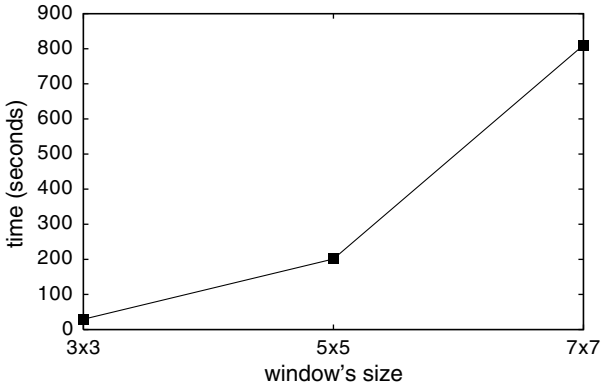


Fig. 8. Execution time for nonparametric registration with *SKP* as a function of the width of the windows used to measure local similarity. Results are shown for an image of 128×128 pixels. For every window's size 200 iterations of the gradient descent were run in every level of the Gaussian pyramid. The tests were run on a pentium 4, 3.0 GHz, PC.

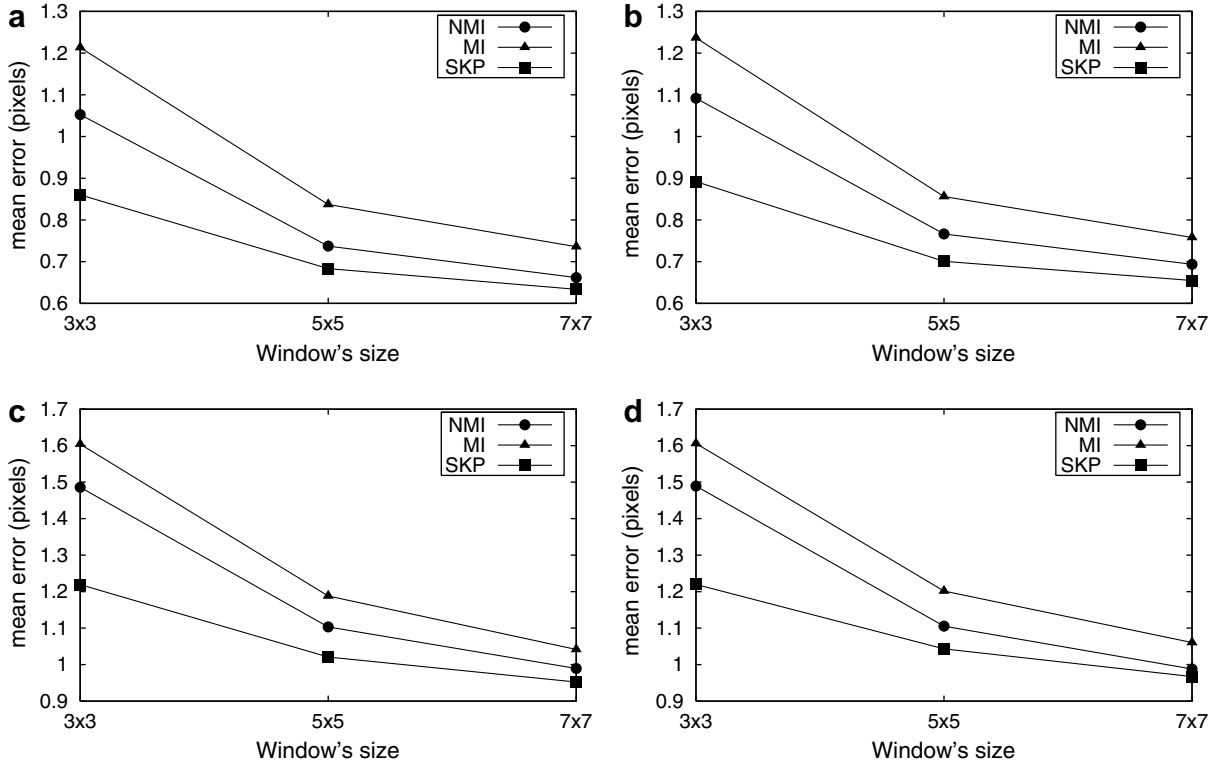


Fig. 7. Mean error in nonparametric registration for different window sizes. The first row shows results for image 6(a) and reference images generated using the tone transfer function $f_1(I) = 100(\frac{I}{100})^{1.35}$ (left plot), and $f_2(I) = 100(1 - \frac{I}{100})^{1.35}$ (right plot). The second row shows the corresponding results for image 6(b).

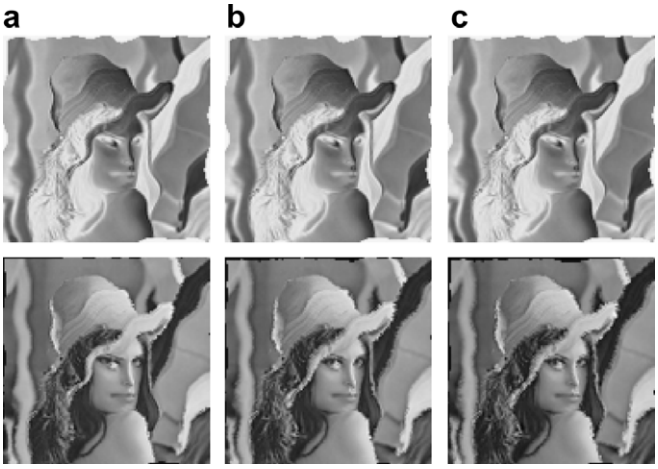


Fig. 9. Registered images for a specific transformation. In the first row, the reference image is shown (the same for each case), and in the second the registered images for *SKP* (left), *NMI* (center) and *MI* (right). The estimation of the deformation field was done locally using windows of 3×3 pixels around every pixel in the images. The respective errors were: 1.23, 1.57 and 1.60 pixels.

tered images by the three methods for a specific transformation are shown in Fig. 9.

5. Conclusions

In this paper, we have proposed the use of a new similarity measure for image registration, based on a novel concept of kernel-predictability for random variables. The performance of our registration method was compared with mutual information and normalized mutual information in different registration situations, including nonparametric registration, and we have shown experimentally that using our method, important reductions in registration errors are obtained, mainly when used for large transformations and in situations where only a small sampling is available. This robustness is due to the fact that the new similarity measure is controlled by the most important features in the images.

Appendix A.

An upper bound for the registration measure *SKP* for the discrete case may be found in the following way: suppose one uses a kernel K to measure *KP* for the intensity distributions of a pair of images I, J, which has the property: $K(i,i) = 1 \geqslant K(i,j)$, for $i \neq j$. One may then construct a separable kernel K_2 for measuring *KP* for the joint distribution $p_{IJ}(I,J)$ as:

$$K_2((i_1,j_1),(i_2,j_2)) = K(i_1,i_2)K(j_1,j_2)$$

we now have:

$$\begin{aligned} KP(p_{IJ}(I,J)) &= \sum_{i_1}\sum_{i_2}\sum_{j_1}\sum_{j_2} p_{IJ}(i_1,j_1)p_{IJ}(i_2,j_2) \\ &\quad \times K_2((i_1,j_1),(i_2,j_2)) \\ &= \sum_{i_1}\sum_{i_2} p_I(i_1)p_I(i_2)K(i_1,i_2) \\ &\quad \times \sum_{j_1}\sum_{j_2} p(j_1|i_1)p(j_2|i_2)K(j_1,j_2) \\ &\leqslant \sum_{i_1}\sum_{i_2} p_I(i_1)p_I(i_2)K(i_1,i_2) = KP(p(I)) \end{aligned}$$

In a similar way, one can see that $KP(p_{IJ}(I,J)) \leqslant KP(p(J))$, so that $SKP(I,J) \leqslant \frac{1}{2}$.

Now, consider a reference image I_R and a transformed image I_T, and assume that when the transformation T^*, which correctly aligns both images, is used, one has that the intensities i_R, i_{T^*} are related by a deterministic, invertible tone transfer function Φ, so that $p(i_{T^*}|i_R) = \delta(i_{T^*} - \Phi(i_R))$. Assume also that $K(i,j) = \delta(i-j)$ (a Kronecker delta function). In this case, from the above equation one can see that $KP(p(I_R, I_{T^*})) = KP(p(I_R)) = KP(p(I_{T^*}))$, so that $SKP(I_R, I_{T^*}) = \frac{1}{2}$, which means that *SKP* reaches its global maximum when $T = T^*$.

References

[1] http://www.bic.mni.mcgill.ca/brainweb/.
[2] G. Aubert, R. Deriche, P. Kornprobst, Computing optical flow via variational techniques, SIAM Journal on Applied Mathematics 60 (1) (2000) 156–182.
[3] A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, G. Marchal, Automated multi-modality image registration based on information theory, Information Processing in Medical Imaging (1995) 263–274.
[4] E. D'Agostino, F. Maes, D. Vandermeulen, P. Suetens, A viscous fluid model for multimodal image registration using mutual information, MICCAI (2002) 541–548.
[5] N. Dowson, R. Bowden, Metric mixtures for mutual information tracking, ICPR 2 (2004) 752–756.
[6] R.O. Duda, P.E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, 1973.
[7] E. Geoffrey, Mutual information as a stereo correspondence measure, Technical Report MS-CIS-00-20, University of Pennsylvania, 2000.
[8] L. Gottesfeld, A survey of image registration techniques, ACM Computing Surveys 24 (4) (1992) 325–376.
[9] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning, Springer Verlag, New York, 2003.
[10] G. Hermosillo, C. Chefd'hotel, O. Faugeras, Variational methods for multimodal image matching, International Journal of Computer Vision 50 (3) (2002) 329–343.
[11] B. Horn, B.G. Schunck, Determining optical flow, Artificial Intelligence 17 (1981) 185–203.
[12] P.W. Josien, J.B. Pluim, A. Maintz, A. Viergever, Mutual information based registration of medical images: a survey, IEEE Transactions on Medical Imaging 22 (8) (2003) 986–1004.
[13] J. Kim, V. Kolmogorov, R. Zabih, Visual correspondence using energy minimization and mutual information, ICCV (2003) 1033–1040.
[14] A.J. Lee, U-Statistics, Theory and Practice, Marcel Dekker Inc., New York, 1990.
[15] E. Lehmann, Elements of Large Sample Theory, Springer Verlag, New York, 1999.

[16] B. Lucas, T. Kanade, An iterative image registration technique with an application to stereo vision, IJCAI81 (1981) 674–679.
[17] A. Maintz, M.A. Viergever, A survey of medical image registration, Medical Image Analysis 2 (1) (1998) 1–36.
[18] W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, Numerical Recipes in C: The Art of Scientific Computing, Cambridge University Press, Cambridge, 1999.
[19] M. Singh, H. Arora, N. Ahuja, Robust registration and tracking using kernel density correlation, CVPRW (2004) 174.
[20] C. Studholme, D.L.G. Hill, D.J. Hawkes, An overlap invariant entropy measure of 3d medical image alignment, Pattern Recognition 32 (1) (1999) 71–86.
[21] R. Szeliski, J. Coughlan, Spline-based image registration, International Journal of Computer Vision 22 (3) (1997) 199–218.
[22] J.-P. Thirion, Image matching as a diffusion process: an analogy with Maxwell's demons, Medical Image Analysis 2 (3) (1998) 243–260.
[23] P. Viola, W. Wells III, Alignment by maximization of mutual information, ICCV (1995) 16–23.
[24] C. Yang, R. Duraiswami, L. Davis, Efficient mean-shift tracking via a new similarity measure, CVPR (2005) 176–183.
[25] S.K. Zhou, R. Chellappa, Probabilistic identity characterization for face recognition, CVPR (2004) 805–812.
[26] Y.M. Zhu, S.M. Cochoff, Influence of implementation parameters on registration of mr and spect brain images by maximization of mutual information, Journal of Nuclear Medicine 43 (2) (2002) 160–166.

Please cite this article in press as: H.F. Gómez-García et al., Image registration based on kernel-predictability, Comput. Vis. Image Understand. (2008), doi:10.1016/j.cviu.2008.02.001

BIBLIOGRAFíA

[ADK99] G. Aubert, R Deriche, and P Kornprobst. Computing optical flow via variational techniques. *SIAM Journal of Applied Mathematics*, 60(1):156–182, 1999.

[BDC+93] Stephen L. Bacharach, Margaret A. Douglas, Richard E. Carson, Paul J. Kalkowski, Nanette M.T. Freedman, Pasquale Perrone-Filardi, and Robert 0. Bonow. Three-dimensional registration of cardiac positron emission tomography attenuation scans. *The Journal of Nuclear Medicine*, 34:311–321, 1993.

[BFB04] A. Bardera, M. Feixas, and I. Boada. Normalized similarity measures for medical image registration. In *SPIE, International Symposium in Medical Imaging*, pages 0–0, 2004.

[BGL+93] Valentino Bettinardi, Maria Carla Gilardi, Giovanni Lucignani, Claudio Landoni, Giovanna Rizzo, Giuseppe Striano, and Ferruccio Fazio. A procedure for patient repositioning and compensation for misalignment between transmission and emission data in pet heart studies. *The Journal of Nuclear Medicine*, 34:137–142, 1993.

[BHS97] Alireza Bab-Hadiashar and David Suter. Optic flow calculation

using robust statistics. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, pages 988–993, 1997.

[Bis06] Chirstopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer Verlag, 2006.

[BR96] Michael J. Black and Anand Rangarajan. On the unification of line processes, outlier rejection, and robust statistics with aplplications in early vision. *International Journal of Computer Vision*, 19(1):57–92, 1996.

[Bra] http://www.bic.mni.mcgill.ca/brainweb/.

[BT01] Torsten Butz and Jean-Philippe Thiran. Affine registration with feature space mutual information. In W.Ñiessen and M. Viergever, editors, *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2001)*, pages 549–556, 2001.

[CCY+03] Ho-Ming Chan, Albert C.S. Chung, Simon C.H. Yu, Alexander Norbash, and William M. Wells III. Multi-modal image registration by minimizing kullback-leibler distance between expected and observed joint class histograms. In *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR03)*, page 570, 2003.

[CDD94] Qin-Sheng Chen, Michel Defrise, and F. Deconinck. Symmetric phase-only matched filtering of fourier-mellin transforms for im-

age registration and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(12):1156–1168, 1994.

[CMD+95a] A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, and G. Marchal. Automated multi-modality image registration based on information theory. In Y. Bizais, C. Barillot, and R. Di Paola, editors, *Information Processing in Medical Imaging*, pages 263–274, 1995.

[CMD+95b] A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, and G. Marchal. Automated multi-modality image registration based on information theory. In *Information Processing in Medical Imaging*, pages 263–274, 1995.

[CV01] Tony F. Chan and Luminita Vese. A level set algorithm for minimizing the mumford-shah functional in image processing. In *IEEE Workshop on Variational and Level Set Methods (VLSM'01)*, page 161, 2001.

[CWFT05] Yunqiang Chen, Hongcheng Wang, Tong Fang, and Jason Tyan. Mutual information regularized bayesian framework for multiple image restoration. In *Tenth IEEE International Conference on Computer Vision (ICCV)*, pages 190–197, 2005.

[DB08] Nicholas Dowson and Richard Bowden. Mutual information for lucas-kanade tracking (milk): An inverse compositional formulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(1):180–185, 2008.

[DH73] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* John Wiley and Sons., 1973.

[DKB08] Nicholas Dowson, Timor Kadir, and Richard Bowden. Estimating the joint statistics of images using non-parametric windows with application to registration using mutual information. *(por publicarse en)IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[EP08] Georgios D. Evangelidis and Emmanouil Z. Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence,*, http://doi.ieeecomputersociety.org/10.1109/TPAMI.2008.113, 2008.

[FMST07] De Falco, D. Maisto, U. Scafuri, and E. Tarantino. Distributed differential evolution for the registration of remotely sensed images. In *15th EUROMICRO International Conference on Parallel, Distributed and Network-Based Processing (PDP'07)*, pages 358–362, 2007.

[Gar02] Héctor Fernando Gómez García. Estrategias evolutivas aplicadas al problema de registro de imágenes. Master's thesis, Centro de Investigación en Matemáticas (CIMAT), Agosto 2002.

[GG84] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE*

Transactions on Pattern Analysis and Machine Intelligence, 6:721–741, 1984.

[Got92] Lisa Gottesfeld. A survey of image registration techniques. *ACM Computing Surveys*, 24(4):325–376, 1992.

[GVA+02] H. F. Gómez, A. G. Vega, A. H. Aguirre, J. L. Marroquín, and C. A. C. Coello. Robust multiscale affine 2d-image registration through evolution strategies. In *LNCS 2439. Parallel Problem Solving From Nature-PPSN VII*, pages 740–748, 2002.

[GXLP05] Xiaoxin Guo, Zhiwen Xu, Yinan Lu, and Yunjie Pang. An application of fourier-mellin transform in image registration. In *Proceedings of the 2005 The Fifth International Conference on Computer and Information Technology (CIT05)*, pages 619–623, 2005.

[HBC+03] P. Hellier, C. Barillot, I. Corouge, B. Gibaud, G. Le Goualher, D.L. Collins, A. Evans, G. Malandain, N. Ayache, G.E. Christensen, and H.J. Johnson. Retrospective evaluation of intersubject brain registration. *IEEE Transactions on Medical Imaging*, 22(9):1120–1130, 2003.

[Hog03] William Scott Hoge. A subspace identification extension to the phase correlation method. *IEEE Trans. Medical Imaging*, 22(2):277–280, 2003.

[HR81] Berthold K.P. Horn and Brian G. Rhunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981.

[Jan02] Bernd Jane. *Digital Image Processing.* Springer Verlag, 2002.

[JK03] Ramin Zabih Junhwan Kim, Vladimir Kolmogorov. Visual correspondence using energy minimization and mutual information. In *Ninth IEEE International Conference on Computer Vision (ICCV'03)*, page 1033, 2003.

[JMH+90] Larry Junck, John 0. Moen, Gary D. Hutchins, Morton B. Brown, and David E. Kuhi. Correlation methods for the centering, rotation, and alignment of functional brain images. *The Journal of Nuclear Medicine*, 31(7):1220–1226, 1990.

[JPMV03] P. W. Josien, J. B. Pluim, A. Maintz, and A. Viergever. Mutual information based registration of medical images: a survey. *IEEE Transactions on Medical Imaging*, 22(8):986–1004, 2003.

[KAM04] Yosi Keller, Amir Averbuch, and Ofer Miller. Robust phase correlation. In *Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)*, pages 740–743, 2004.

[KK06] Yeon-Ho Kim and Avinash C. Kak. Error analysis of robust optical flow estimation by least median of squares methods for the varying illumination model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1418–1435, 2006.

[KSA05] Yosi Keller, Yoel Shkolnisky, and Amir Averbuch. The angular difference function and its application to image registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):969–976, 2005.

[Lee90] A. J. Lee. *U-Statistics, Theory and Practice*. Marcel Dekker Inc. New York, 1990.

[Leh99] E.L. Lehmann. *Elements of Large Sample Theory*. Springer Verlag, New York, 1999.

[Li01] Stan Z. Li. *markov Random Field in Image Analysis*. Springer Verlag, 2001.

[LP01] B. Likar and F. Pernus. A hierarchical approach to elastic registration based on mutual information. *Image and Vision Computing*, 19(1–2):33–44, 2001.

[MMP87] J. Marroquin, S. Mitter, and T. Poggio. Probabilistic solution of ill-posed problems in computational vision. *J. Am. Statistical Assoc.*, 82:76–89, 1987.

[MSB03] Jose L. Marroquin, Edgar Arce Santana, and Salvador Botello. Hidden markov measure field models for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(11):1380–1387, 2003.

[MV98] Antoine. Maintz and M. A. Viergever. A survey of medical image registration. *Medical Image Analysis*, 2(1):1–36, 1998.

[MVS99] Frederik Maes, Dirk Vandermeulen, and Paul Suetens. Comparative evaluation of multiresolution optimization strategies for multimodality image registration by maximization of mutual information. *Medical Image Analysis*, 3(4):373–386, 1999.

[Neg98] Shahriar Negahdaripour. Revised definition of optical flow: Integration of radiometric and geometric cues for dynamic scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(9):961–979, 1998.

[PD99] Nikos Paragios and Rachid Deriche. Geodesic active regions for supervised texture segmentation. In *Seventh International Conference on Computer Vision (ICCV'99) - Volume 2*, page 926, 1999.

[PQC08] Wei Pan, Kaihuai Qin, and Yao Chen. An adaptable-multilayer fractional fourier transform approach for image registration. *(por publicarse en)IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[PTVP99] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and Flannery B. P. *Numerical Recipes in C: The Art of Scientific Computing.* Cambridge University Press, 1999.

[RBR06] Ajit Rajwade, Arunava Banerjee, and Anand Rangarajan. A new method of probability density estimation with application to mutual information based image registration. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR06)*, pages 1769–1776, 2006.

[RBR08] Ajit Rajwade, Arunava Banerjee, and Anand Rangarajan. Probability density estimation using isocontours and isosurfaces: Ap-

plication to information theoretic image registration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2008.

[RCL98] C. E. Rodriguez-Carranza and M. H. Loew. A weighted and deterministic entropy measure for image registration using mutual information. In Ed. Bellingham K. M. Hanson, editor, *Medical Imaging: Image Processing*, pages 155–166, 1998.

[RMPA98] A. Roche, G. Malandain, X. Pennec, and N Ayache. The correlation ratio as a new similarity measure for multimodal image registration. In W. Wells, A. Colchester, and S. Delp, editors, *Medical Image Computing and Computer-Assisted Intervention-MICCAI98,*, pages 1115–1124, 1998.

[ROM05] Mariano Rivera, Omar Ocegueda, and José L. Marroquín. Entropy controlled gauss-markov random measure field for early vision. In *LNCS 3752*, pages 137–148, 2005.

[ROM07] M. Rivera, O. Ocegueda, and J.L Marroquin. Entropy-controlled quadratic markov measure field models for efficient image segmentation. *IEEE Transactions on Image Processing*, 16(12):3047–3057, 2007.

[SAA04] Maneesh Singh, Himanshu Arora, and Narendra Ahuja. Robust registration and tracking using kernel density correlation. In *CVPRW*, page 174, 2004.

[SD06] Vinay Sharma and James W. Davis. Feature-level fusion for object segmentation using mutual information. In *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, page 139, 2006.

[SG07] Shaoyan Sun and Chonghui Guo. Image registration by minimizing tsallis divergence measure. In *Fourth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2007)*, pages 712–715, 2007.

[Sha48] C. E. Shannon. A mathematical theory of communication. Technical report, Bell Syst. Tech, 1948.

[SHH99] C. Studholme, D. L. G. Hill, and D. J. Hawkes. An overlap invariant entropy measure of 3d medical image alignment. *Pattern Recognit*, 32(1):71–86, 1999.

[SLP04] Darko Skerl, Bostjan Likar, and Franjo Pernus. Evaluation of nine similarity measures used in rigid registration. In *17th International Conference on Pattern Recognition (ICPR'04)*, pages 794–797, 2004.

[SOC99] Harold Stone, Michael Orchard, and Ee-Chien Chang. Subpixel registration of images. In *33rd Asilomar Conference on Signal, Systems and Computers*, 1999.

[TLCH02] Chin-Hung Teng, Shang-Hong LaI, Yung-Sheng Chen, and Wen-Hsing Hsu. Robust computation of optical flow under non-uniform ulumination variations. In *Proceedings of the 16 th*

International Conference on Pattern Recognition, page 10327, 2002.

[TU00] Philippe Thvenaz and Michael Unser. Optimization of mutual information for multiresolution image registration. *IEEE Transactions on Image Processing*, 9(12):2083–2099, 2000.

[VSOB99] Klaus Voss, Herbert Suesse, Wolfgang Ortmann, and Torsten Baumbach. Shift detection by restoration. *Pattern Recognition*, 32(12):2067–2068, 1999.

[VWI95] P. Viola and W. Wells III. Alignment by maximization of mutual information. In *ICCV*, pages 16–23, 1995.

[WCLY06] Wei Wang, Houjin Chen, Jupeng Li, and Jiangbo Yu. A registration method of fundus images based on edge detection and phase-correlation. In *Proceedings of the First International Conference on Innovative Computing, Information and Control (ICICIC'06)*, pages 572–576, 2006.

[YC08] Qiyao Yu and David Clausi. Irgs: Image segmentation using edge penalties and region growing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, http://doi.ieeecomputersociety.org/10.1109/TPAMI.2008.15, 2008.

[YDD05] Changjiang Yang, Ramani Duraiswami, and Larry Davis. Efficient mean-shift tracking via a new similarity measure. In *CVPR*, pages 176–183, 2005.

[YMLL07] Anrong Yang, Lingqi Meng, Jianzhen Luo, and Caixing Lin1. A rapid registration framework for medical images. In *Fourth International Conference on Image and Graphics (ICIG 2007)*, pages 731–736, 2007.

[ZC02] Y. M. Zhu and S. M. Cochoff. Influence of implementation parameters on registration of mr and spect brain images by maximization of mutual information. *J. Nucl. Med.*, 43(2):160–166, 2002.

[ZC04] Shaohua Kevin Zhou and Rama Chellappa. Probabilistic identity characterization for face recognition. In *CVPR*, pages 805–812, 2004.

[ZF03] Barbara Zitova and Jan Flusser. Image registration methods: a survey. *Image and Vision Computing*, 21:977–1000, 2003.

[ZR04] Jie Zhang and Anand Rangarajan. Affine image registration using a new information metric. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR04)*, pages 848–855, 2004.

[ZY96] Song Chun Zhu and Alan Yuille. Region competition: Unifying snakes, region growing, and bayes/mdl for multiband image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(9):884–900, 1996.

[ZZSZ05] Hongying Zhang, Xiaozhou Zhou, Jizhou Sun, and Jiawan Zhang. A novel medical image registration method based on

mutual information and genetic algorithm. In *International Conference on Computer Graphics, Imaging and Visualization (CGIV'05)*, pages 221–226, 2005.