Clasificación de configuraciones de manos del Lenguaje de Señas Argentino con ProbSOM

Franco Ronchetti, Facundo Quiroga, César Estrebou, Laura Lanzarini

Instituto de Investigación en Informática LIDI, Facultad de informática,
Universidad Nacional de La Plata
{fronchetti,fquiroga,cesarest,laural}@lidi.unlp.edu.ar

Resumen El reconocimiento automático de lenguaje de señas es una temática actual de sumo interés dentro del reconocimiento de gestos humanos. Por un lado, su complejo campo de aplicación presenta un desafío que requiere la intervención de diferentes áreas del conocimiento como el procesamiento de video, de imágenes, los sistemas inteligentes y la lingüística. Por otro lado, la correcta clasificación de las señas podría facilitar la traducción e integración a personas con discapacidad auditiva. El presente trabajo tiene dos principales aportes: por un lado la confección de una base de datos de configuraciones de manos del Lenguaje de Señas Argentino (LSA), temática prácticamente no encontrada en el estado del arte. En segundo lugar, el procesamiento de las imágenes, extracción de descriptores y posterior clasificación de la configuración por medio de una adaptación supervisada de los mapas auto-organizativos llamada ProbSom. Dicha técnica se compara con otras del estado del arte como Máquinas de Soporte Vectorial (SVM), Random Forest, y Feedforward Neural Networks.
La base de datos desarrollada contiene 800 imágenes con 16 configuraciones de LSA lo que permite ser un paso inicial hacia la confección de una base de datos de señas argentinas completa. A su vez, la extracción de características propuestas sumadas al clasificador neuronal demostraron ser sumamente eficaces, con una tasa de acierto superior al 90 %.

Keywords: reonocimiento de configuraciones de manos, reconocimiento de formas de mano, reconocimiento de lenguajes de seña, ProbSom, mapas auto organizativos, SOM, transformada Radon, SIFT, Scale-Invariant Feature Transform

1. Introducción

El reconocimiento automático de señas es un problema multidisciplinar sumamente complejo que hoy en día sigue sin ser resuelto en forma total. Si bien en el último tiempo han habido avances en el reconocimiento de gestos, impulsados principalmente por el desarrollo de nuevas tecnologías, aún queda un largo camino por recorrer para construir aplicaciones precisas y robustas que permitan la traducción e interpretación de los gestos realizados por un intérprete [1]. La

compleja naturaleza de los gestos motivan esfuerzos de diversas áreas de investigación como interacción hombre-máquina, visión por computador, análisis de movimientos, aprendizaje automático y reconocimiento de patrones. El lenguaje de señas, y particularmente el Lenguaje de Señas Argentino (LSA), es una temática muy impulsada actualmente por gobiernos y universidades para incluir a persona hipoacúsicas en diferentes entornos. Existe poca documentación y aún menos información en formato digital.

La tarea completa de reconocer un gesto de lenguaje de señas involucra diferentes pasos: la ubicación de las manos del intérprete, el reconocimiento de las formas de las manos (configuraciones), y el seguimiento de las manos para detectar el movimiento realizado, interpretación semántica y traducción al lenguaje escrito [1]. Estas tareas pueden ser desarrolladas y evaluadas en forma separada ya que cada una tiene su complejidad particular. Existen diferentes enfoques para el seguimiento de la mano: algunos utilizando sistemas 3D como el MS Kinect y otros simplemente con una imagen 2D proveniente de una cámara RGB. Incluso existen sistemas con sensores de movimiento como guantes especiales, acelerómetros, etc.

El trabajo presentado en este documento se enfoca en el problema de clasificación de configuraciones de manos. En particular, este se centra en la extracción de características representativas de la mano y en el reconocimiento de dichas configuraciones utilizando una variante de red neuronal competitiva supervisada denominada ProbSom [3]. El trabajo tiene como finalidad generar una subunidad (*Handshape Sub-Unit*) de procesamiento para el reconocimiento automático de lenguaje de señas. En [5] se incorpora el concepto de *subunidad léxica* para modularizar el reconocimiento del gesto.

Una particularidad del lenguaje de señas es que cada región a nivel mundial tiene su propio léxico y grupo de señas que lo representan. Esto lo hace un problema diverso, y diferente de abordar en cada región, ya que nuevos gestos o configuraciones de manos involucran nuevos desafíos no contemplados con anterioridad. En particular, para el Lenguajes de Señas Argentino (LSA) practicamente no existen sistemas y bases de datos que representen los gestos que posee. En este trabajo se aborda también la confección de una base de datos de 16 configuraciones de LSA interpretados por 10 personas distintas. Las imágenes obtenidas fueron utilizadas luego para el proceso de extracción de características y posterior clasificación.

En la literatura existen numerosos trabajos desarrollados que abordan el reconocimiento automático de lenguajes de señas. No obstante, cada trabajo presenta un escenario particular, a veces difícil de replicar completamente, o con ciertas limitaciones. Por ejemplo, diferente trabajos utilizan sensores de profundidad como el MS Kinect, o similares para capturar imágenes 3D. En [8],[12] y [9] se utilizan imágenes de profundidad para clasificar configuraciones del lenguaje de señas norteamericano (ASL). Estos enfoques en general presentan dos problemas: por un lado la necesidad de contar con un equipo de similares características con el que fue probado, y por otro lado la alta tasa de error que todavía tienen estos dispositivos (al menos los de un costo bajo) para calcular

las imágenes de profundidad. Otros enfoques, como el que se presenta en este trabajo, utilizan sólo imágenes RGB. En [10] se crea un modelo probabilístico de color de piel para detectar y seguir las manos del intérprete en un video. En [2] se utiliza este modelo para segmentar las manos y aplicar un clasificador basado en Modelos de Markov. En general los sistemas basados únicamente en color de piel no son robustos a la variabilidad en el fondo o la vestimenta del intérprete, y en las oclusiones mano-mano o mano-cara. Para realizar un reconocimiento de la posición de la mano suele ser necesario adicionar información morfológica al filtrado de color. Por último, en [1] se hace una gran revisión del estado del arte en el reconocimiento de lenguaje de señas.

El presente documento se organiza de la siguiente manera: en la sección 2 se describe la base de datos generada, el procesamiento de las imágenes y la extracción de características de la mano y el modelo de clasificación utilizado. En la sección 3 se detalla la experimentación y finalmente en la sección 4 se exponen las conclusiones generales.

2. Métodos

2.1. Base de datos de configuraciones de Lengua de Señas Argentina (LSA16)

La base de datos de configuraciones de Lengua de Señas Argentina [1], creada con el propósito de producir un diccionario de LSA y entrenar un traductor automático de señas, contiene 800 imágenes en donde 10 sujetos realizaron 5 repeticiones de 16 tipos distintos de configuraciones de mano utilizadas en distintas señas de dicho lenguaje. Las configuraciones fueron elegidas dentro de las más utilizadas en el léxico, y se pueden observar en la figura 1. Cada configuración fue realizada repetidamente en diferentes posiciones y diferentes rotaciones en el plano perpendicular a la cámara, para generar mayor diversidad y realismo en la base de datos.

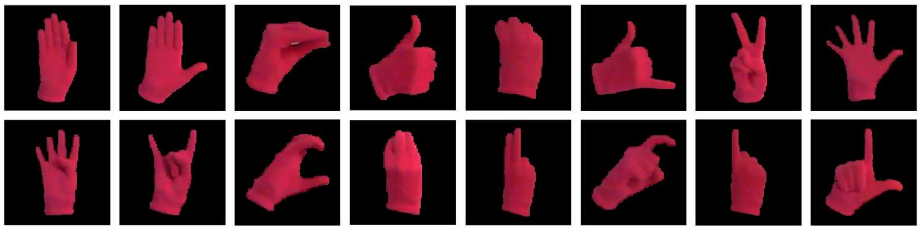

Figura 1. Ejemplos de cada clase de la base de datos LSA16

[1] Se puede encontrar más información sobre esta base de datos en http://facundoq.github.io/unlp/lsa16/.

Los sujetos vistieron ropa negra, sobre un fondo blanco con iluminación controlada, como se observa en la figura 2. Para la simplificar el problema de segmentación de la mano dentro de una imagen, los sujetos utilizaron guantes de tela con colores fluorescentes en sus manos. Esto resuelve parcialmente pero de un modo muy eficaz el reconocimiento de la posición de la mano y carece de los problemas existentes en los modelos de piel. Por otro lado, propone un artefacto simple y económico al momento de realizar pruebas o confeccionar una aplicación real.

Figura 2. Imágenes no segmentadas de la base de datos LSA16

2.2. Preprocesamiento y Descriptores

A continuación se detalla el preprocesamiento realizado de las imágenes de manos segmentadas, los descriptores calculados en base a la imagen preprocesada y el contorno de la mano calculado en base a la misma, y el modelo de clasificación presentado. La entrada a la etapa de preprocesamiento consiste de una imagen donde los únicos pixeles no negros corresponden a la mano.

Preprocesamiento Para cada imagen, y en base a la componente conexa única determinada por la máscara de segmentación de la mano, se calculan los ejes principales de los pixeles de la mano y con ellos la inclinación ϕ de la misma. Luego, se rota la imagen $-\phi^o$ para llevarla a una orientación canónica. Como esta orientación es insensible a rotaciones de 180^o de la mano, puede que la imagen quede orientada hacia arriba o abajo. Para corregir esto, se calcula la cantidad de cruces de cada línea horizontal posible en la imagen, y se estima la posición de los dedos en base si la moda de la cantidad de cruces se encuentra en la parte superior o inferior de la imagen.

La imagen se re-samplea sin afectar su relación de aspecto a una tamaño de 128×128 y se re-posiciona de modo de que la misma quede centrada. El contorno de la mano se obtiene aplicando un filtro de bordes a la máscara de segmentación de la mano, la cual contiene una sola componente conexa.

Descriptores A continuación se describen dos descriptores, uno basado en la transformada de Radon, y otro en los Scale-Invariant Feature Transform (SIFT).

Figura 3. De izquierda a derecha: Imagen segmentada, imagen orientada, imagen con rotación corregida, máscara de segmentación y contorno.

Transformada de Radon La transformada de Radon ha sido utilizada en el pasado para reconocer objetos y también para identificar a personas en base a las características de su mano [4].

La transformada de Radon de una imagen 2D $f : \mathbb{R}^2 \to \mathbb{R}$ se define como una integral de línea sobre la imagen. La línea L a través de la cual se integra está dada por un par (b, θ), donde b es distancia al origen de la línea y θ el ángulo con el eje horizontal de la imagen. Está dada por la fórmula:

$$R_{(b,\theta)} = \iint_{L_{(b,\theta)}} f(\boldsymbol{x})|d\boldsymbol{x}| = \int_{-\infty}^{\infty} f(x(t), y(t))dt$$
$$= \int_{-\infty}^{\infty} f(tsin\theta + bcos\theta, -tcos\theta + bsin\theta)dt$$

Aplicando la versión discreta de la misma a la imagen segmentada para todas las combinaciones de valores enteros de (b, θ) posibles (1.,180 para θ, un valor K dependiente del tamaño de la imagen para b), obtenemos un descriptor $R \in \mathbb{R}^{180 \times K}$. Luego para reducir la dimensionalidad r se re-samplea a un tamaño fijo $r \in \mathbb{R}^{32 \times 32}$. Este descriptor se utiliza como global considerándolo un vector en $r' \in \mathbb{R}^{32^2}$, o como 32 descriptores locales tomando cada fila $r_i,\ i = 1, \ldots, 32,\ r_i \in \mathbb{R}^{32}$ como un descriptor local. Cada r_i entonces contiene una aproximación suave a los $R_{(b,\theta)}$ para todo b, y donde θ corresponde aproximadamente la media de un subconjunto de ángulos contiguos.

En particular, como el clasificador que se presenta, el ProbSom, tiene como entrada un conjunto de cardinalidad arbitraria de vectores, se utilizaron los vectores r_i para el mismo, y el vector completo r' para el resto de los clasificadores probados.

SIFT Un descriptor SIFT es un histograma espacial 3D de los gradientes de una imagen, que caracteriza la apariencia de un punto de interés. Para ello, con el gradiente de cada pixel se calcula un descriptor más elemental formado por la ubicación del pixel y la orientación del gradiente. Dado un posible punto de interés, estos descriptores elementales son pesados por la norma del gradiente y acumulados en un histograma 3D que representa el descriptor SIFT de la región alrededor del punto de interés. Al formar el histograma, se le aplica a

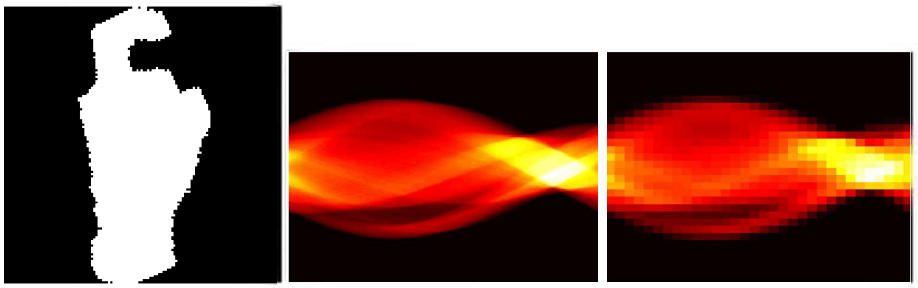

Figura 4. Imagen original, transformada de Radon, transformada de Radon resampleada

los descriptores elementales una función de peso gaussiana para darle menos importancia a los gradientes que están más lejos del centro punto de interés.

Los descriptores SIFT han sido aplicados a varias tareas de visión por computadoras, incluyendo el reconocimiento de configuraciones de mano [13] y reconocimiento de rostros [7].

2.3. Modelo de clasificación ProbSom

ProbSOM [3] es una adaptación probabilística de los mapas auto-organizados de Kohonen(SOM)[6]. Estos mapas son redes competitivas no supervisadas que configuran sus neuronas para representar la distribución de los datos de entrada procesados durante la fase de entrenamiento. Como resultado de esta fase de aprendizaje se obtiene una red donde cada neurona aprende a representar un área del espacio de entrada donde agrupa vectores de datos por su similitud o cercanía.

El proceso de entrenamiento del ProbSOM se realiza de la misma manera que en el algoritmo SOM convencional. ProbSOM agrega una etapa adicional luego del entrenamiento para pesar la proporción de representación de cada neurona. Para ello se repasan todos los patrones de entrada y se agrega a cada una de las neuronas ganadoras información acerca de la clase que representa y en que proporción.

El proceso de reconocimiento también es similar al SOM. El mecanismo de respuesta que decide la identificación de una clase consiste en un sistema probabilista. Como cada vector no permite por si solo la identificación de una clase, una secuencia de vectores es requerida. Cuando un conjunto de vectores de características son introducidos en la red, se obtiene un conjunto de neuronas ganadoras donde cada una representa a varias clases con una proporción determinada. La clase identificada será aquella cuya suma de proporciones sea máxima.

ProbSOM ha demostrado ser un algoritmo robusto para resolver problemas de clasificación [7,3,11] donde las clases se representan por un conjunto de vectores de características, donde dichas clases pueden tener en común vectores muy similares dentro de este conjunto.

3. Resultados

3.1. Metodología y Resultados

A continuación se compara la performance resultante de las pruebas llevadas a cabo con distintos métodos y descriptores. En el caso del ProbSom, se realizaron pruebas con los descriptores SIFT y Radon. Además, para el descriptor basado en Radon, se realizaron pruebas con los modelos estándar del estado del arte Máquinas de Soporte Vectorial (SVM), Random Forest, y Feedforward Neural Networks. [2]. En los casos en que los métodos a comparar se comportan con distinta performance dependiendo de sus parámetros internos, reportamos la mejor.

Método	Performance CV
ProbSom con Radon	92,3(±2,05)
ProbSom con SIFT	88,7(±2,50)
Random Forest con Radon	91,0(±1,91)
SVM con Radon	91,2(±1,69)
Feedforward Neural Net con Radon	78,8(±3,80)

Cuadro 1. Porcentajes de reconocimiento correcto de CV para la base de datos LSA16 utilizando validación cruzada aleatoria.

La medida de performance es el porcentaje de ejemplos reconocidos correctamente sobre el total de cada clase. La tabla 3.1 muestra los resultados obtenidos bajo validación cruzada aleatoria estratificada con $n = 30$ repeticiones independientes, utilizando 90 % de las imágenes para entrenar y 10 % para evaluar. Los resultados muestran una performance comparable del ProbSOM frente a otras técnicas de clasificación. Por otro lado, los descriptores de radón mostraron ser mucho más representativos que los vectores SIFT. Esto puede deberse a que generalmente los descriptores SIFT buscan puntos con información particular, para luego realizar *matching* de imágenes, o describir una situación particular. En las imágenes de LSA16 existen diversos puntos muy similares (como las puntas de los dedos) que resultan comunes a muchas clases, lo que dificultad la utilización de SIFT como se había utilizado en [7] para reconocer rostros, utilizando el mismo modelo de clasificación.

Validación inter-sujeto Utilizando la mejor configuración obtenida (descriptor Radon y ProbSOM) se llevó a cabo una validación cruzada inter-sujeto, dejando un sujeto para testeo y entrenando con el resto. La media de los 10 sujetos con $n = 30$ repeticiones independientes fue de 87,9 %(±4,7 %). Como es de esperar, al dejar un sujeto fuera, la tasa de acierto decae, ya que cada persona realiza las configuraciones de forma particular, con tamaños y apariencia de mano propia

[2] Se realizaron además pruebas con descriptores de Fourier, Banco de filtros de Gabor, Local Binary Patterns (no descriptos en este artículo) con resultados inferiores en casi todos los casos a los presentados.

del individuo. No obstante, el sistema sigue mostrando buenos resultados, dando como posibilidad el reconocimiento correcto de una configuración realizada por un nuevo individuo desconocido por el sistema. La figura 5 muestra los resultados obtenidos para cada individuo de la base de datos.

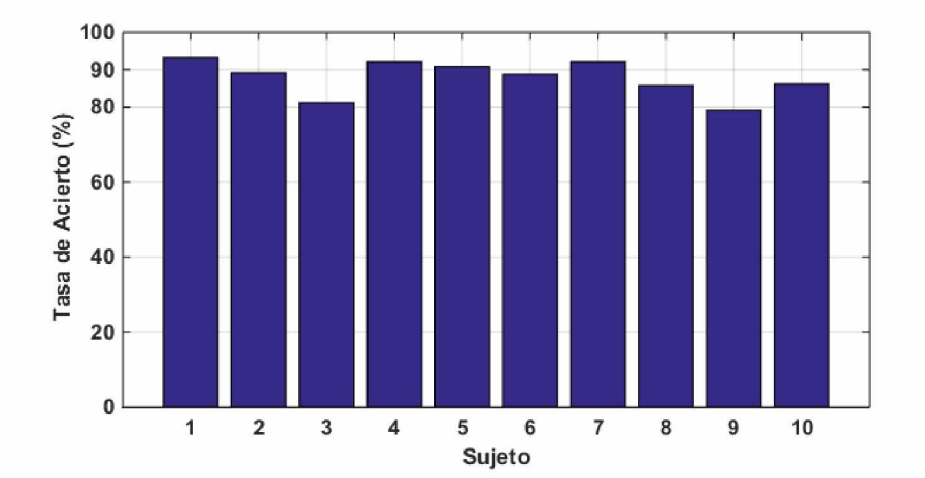

Figura 5. Validación cruzada inter-sujeto para LSA16.

3.2. Discusión

Los descriptores utilizados junto con el modelo de clasificación mostraron ser robustos en la clasificación de las configuraciones de manos en LSA16, incluso con una validación inter-sujeto, dando la posibilidad de incorporar un nuevo individuo desconocido por el sistema. Por otro lado, cabe destacar que la tasa de acierto es similar en todas las clases de la base de datos.

Ya que el ProbSOM funciona de modo probabilístico realizando un ranking de posibles clases candidatas, resulta interesante observar qué ocurre con las imágenes clasificadas erróneamente por el sistema. Si se observa el orden generado por el modelo y la tasa de acierto se obtiene considerando como clasificación correcta tanto a la primer o a la segunda opción, la tasa de acierto general sube de 92,25 % a 96,6 %. Esto demuestra que el modelo, en casi todos los ejemplos de testeo la confusión es entre sólo dos clases. Esto resulta muy interesante si el modelo funciona como un diccionario, ya que podría utilizarse la probabilidad del modelo para mostrar una o dos posibilidades. Del mismo modo, podría volverse a aplicar un clasificador más específico para solucionar la ambigüedad en las situaciones que lo requieran.

4. Conclusión

En este trabajo se presenta una base de datos de configuraciones de manos para el Lenguaje de Señas Argentino (LSA), junto con un modelo de preprocesamiento de las imágenes y clasificación de las configuraciones.

Los resultados de los experimentos de clasificación fueron favorables, mostrando una alta tasa de acierto tanto en la validación aleatoria como en la inter-sujeto. También se llevaron a cabo comparaciones con diferentes descriptores y métodos de clasificación existentes.

El modelo presentado permite la correcta clasificación de las configuraciones de manos, dando la posibilidad de utilizar esto para generar una sub-unidad léxica parte de un descriptor general para una seña de LSA. Se espera también probar la técnica en otras bases de datos existentes en el estado del arte para determinar su aplicabilidad, así como extenderla para utilizar también imágenes de sensores de profundidad.

Referencias

1. Cooper, H., Holt, B., Bowden, R.: Sign language recognition. In: Moeslund, T.B., Hilton, A., Krüger, V., Sigal, L. (eds.) Visual Analysis of Humans: Looking at People, chap. 27, pp. 539 – 562. Springer (Oct 2011), http://www.springer.com/computer/image+processing/book/978-0-85729-996-3
2. Cooper, H., Ong, E.J., Pugeault, N., Bowden, R.: Sign language recognition using sub-units. Journal of Machine Learning Research 13, 2205–2231 (Jul 2012), http://jmlr.csail.mit.edu/papers/volume13/cooper12a/cooper12a.pdf
3. Estrebou, C., Lanzarini, L., Hasperue, W.: Voice recognition based on probabilistic SOM. In: Latinamerican Informatics Conference. CLEI 2010. Paraguay. October 2010. (2010)
4. Gangopadhyay, A., Chatterjee, O., Chatterjee, A.: Hand shape based biometric authentication system using radon transform and collaborative representation based classification. In: Image Information Processing (ICIIP), 2013 IEEE Second International Conference on. pp. 635–639 (Dec 2013)
5. Kadir, T., Bowden, R., Ong, E.J., Zisserman, A.: Minimal training, large lexicon, unconstrained sign language recognition. In: British Machine Vision Conference (2004)
6. Kohonen, T.: Self-organizing formation of topologically correct feature maps. Biological Cybernetics 43(1), 59–69 (1982)
7. Lanzarini, L., Ronchetti, F., Estrebou, C., Lens, L., Fernandez Bariviera, A.: Face recognition based on fuzzy probabilistic SOM. In: IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS), 2013 Joint. pp. 310–314. IEEE (2013)
8. Pugeault, N., Bowden, R.: Spelling it out: Real-time ASL fingerspelling recognition. In: 1st IEEE Workshop on Consumers Depth Cameras for Computer Vision, in conjunction with ICCV'2011 (2011), http://info.ee.surrey.ac.uk/Personal/N.Pugeault/publications/PugeaultBowden2011b.pdf
9. Rioux-Maldague, L., Giguere, P.: Sign language fingerspelling classification from depth and color images using a deep belief network. In: Computer and Robot Vision (CRV), 2014 Canadian Conference on. pp. 92–97. IEEE (2014)

10. Roussos, A., Theodorakis, S., Pitsikalis, V., Maragos, P.: Hand tracking and affine shape-appearance handshape sub-units in continuous sign language recognition. In: Trends and Topics in Computer Vision - ECCV 2010 Workshops, Heraklion, Crete, Greece, September 10-11, 2010, Revised Selected Papers, Part I. pp. 258–272 (2010), http://dx.doi.org/10.1007/978-3-642-35749-7_20
11. Villamonte, A., Quiroga, F., Ronchetti, F., Estrebou, C., Lanzarini, L., Estelrrich, P., Estelrrich, C., Giannechini, R.: A support system for the diagnosis of balance pathologies. In: Congreso Argentino de Ciencias de la Computación. CACIC 2014. Argentina. October 2014. (2014)
12. Zhang, C., Yang, X., Tian, Y.: Histogram of 3d facets: A characteristic descriptor for hand gesture recognition. In: Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on. pp. 1–8. IEEE (2013)
13. Zhu, X., Wong, K.K.: Single-frame hand gesture recognition using color and depth kernel descriptors. In: Pattern Recognition (ICPR), 2012 21st International Conference on. pp. 2989–2992. IEEE (2012)