Object Recognition by Integrating Multiple Image Segmentations

Caroline Pantofaru, Cordelia Schmid, Martial Hebert
The Robotics Institute, Carnegie Mellon University, USA
INRIA Grenoble, LEAR, LJK, France

Gerardo Arellano
Instituto Nacional de Astrofísica Óptica Y Electrónica
Cordinacción de Ciencias Computacionales
Visión Alto Nivel
garellano@ccc.inaoep.mx

8 de junio de 2009

Resumen

1. Unir reconocimiento y segmentación de objetos de una simple imagen no es trivial.
2. No solo requiere de un correcta clasificación, hay que decidir que píxeles pertenecen a un objeto.
3. La segmentación es inestable, afectada por pequeñas perturbaciones de la imagen, selección de características o diferentes algoritmos.
4. Esta inestabilidad nos lleva a usar múltiples segmentaciones de una imagen.
5. ¿Como hacemos la integración de múltiples bottom-up segmentaciones para mejorar la robustez de reconocimiento?

Introducción

1. Segmentación de imágenes en bottom-up es un método posible para proponer un conjunto de píxeles convincente el cual podrían componer un objeto.
2. Desafortunadamente experimentos han demostrado que una región generada por una sola segmentación raramente identifica un objeto. [1] [2]
3. La calidad de la segmentación es altamente variable, depende de la imagen, algoritmos y parámetros (Fig.1).
4. Incluso los humanos no estamos de acuerdo en la partición de una imagen.
5. En un esfuerzo por atacar estas inquietudes se sugiere el uso de múltiples segmentaciones.

Introducción II

1. Se muestra que integrando múltiples segmentaciones proporciona una base más robusta para el reconocimiento y segmentación de un objeto que una sola segmentación.

Dos principios básicos

- Grupo de píxeles el cual están contenidos en la misma región de segmentación de múltiples segmentaciones debería ser uniformemente clasificados.
- El conjunto de regiones generada por las múltiples segmentaciones proporciona características robustas para clasificación de grupos de píxeles

2. Usando múltiples segmentaciones proporciona múltiples oportunidades de descubrir objetos y crear regiones.

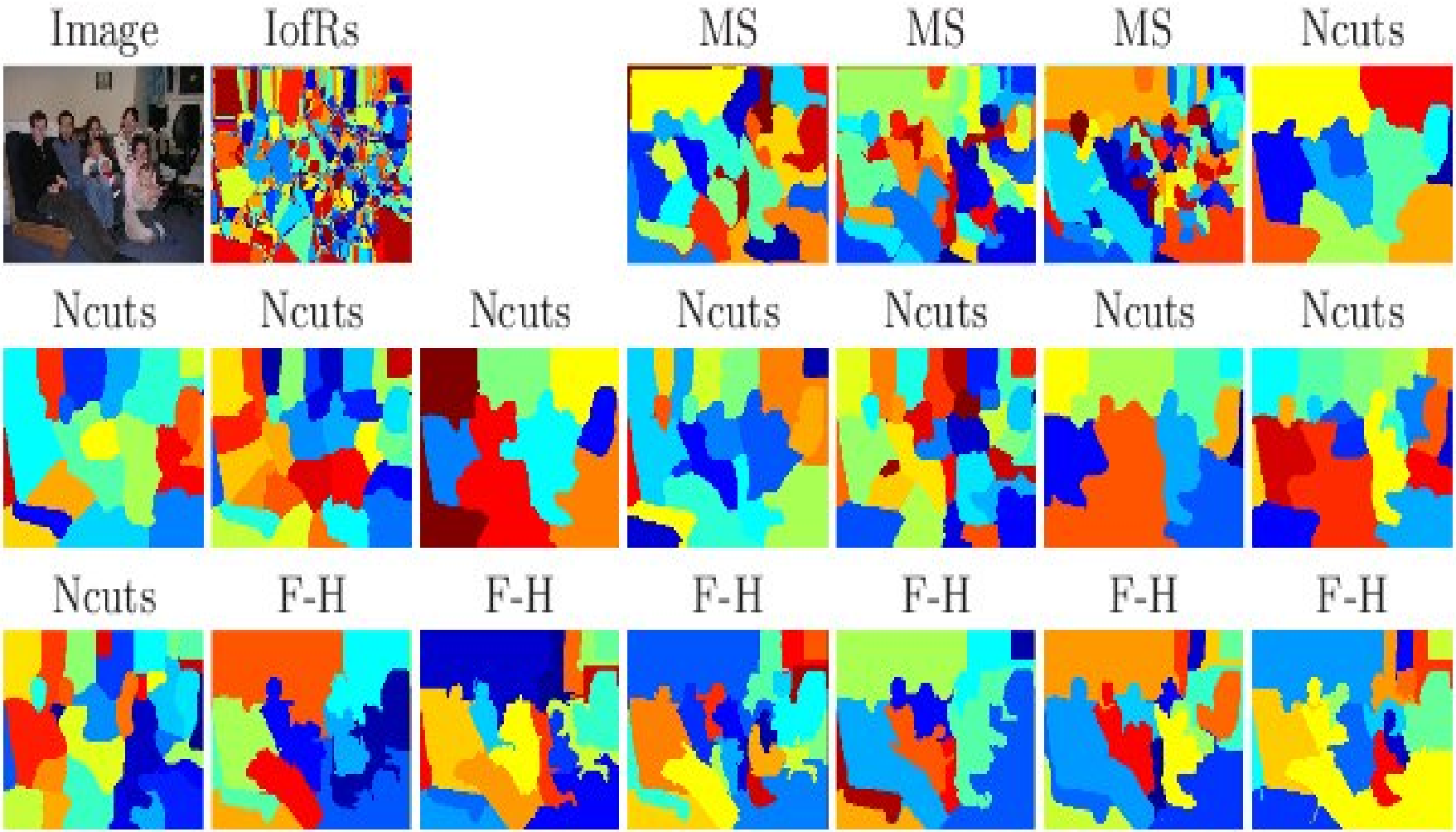

Figura: Ejemplo de intersecciones de regiones (IofRs) de 18 segmentos.

Introducción III

1. Se consideran todas las segmentaciones igualmente útiles.
2. Se intenta aprender la fiabilidad de una región para predecir su contenido.
3. Se intenta ir mas allá de la clasificación de la región independiente modelando regiones adyacentes, utilizando una características de contexto basadas en regiones(RCF).
4. Se consideran el uso de datos adicionales con ruido con supervisión débil.
5. Clasificación de imagen sin localización de objetos y detección de objetos con cajas,
6. El sistema es entrenado usando datos completamente supervisados.

Evaluación

1. 2 conjuntos de datos
2. MSRC 21-clases
3. PASCAL Visual Object Challenge 2007 Segmentation Competition
4. Cada uno contiene múltiples objetos con extremas variaciones; deformaciones, escalas, iluminación, posición, etc.
5. Los resultados son reportados con respecto a un desempeño a nivel de píxel, requiriendo que el objeto exacto sea obtenido.
6. 422 imágenes para entrenar.

Núcleo del Enfoque

Consta de 3 pasos

- Generar múltiples segmentaciones.
- Describir y clasificar cada región.
- Combinar las regiones clasificadas en un mapa de objetos indicando cuales píxeles pertenecen a cada objeto.

Se demuestra que usando una sola segmentación produce resultados de calidad variable, mientras que usando todas las segmentaciones produce mapas de objetos comparables o mejorados.

Núcleo del Enfoque

Generando múltiples segmentaciones

1. Se captura variedad de color, contraste de bordes, textura, tamaño y ruido para producir las múltiples segmentaciones.
2. 18 segmentaciones por imagen.
3. Dimensiones escaladas de 0.5, 0.75 y 1
4. Mean Shift usando posición de píxel, color en el espacio LUV, histograma de textones cuantificado.
5. Ncuts con probabilidad de los limites, generando 9, 21 y 33 regiones.
6. Felzenswalb y Huttenlocher(F-H) con parámetros k=200,500 afectando la escala de las regiones finales.

Núcleo del Enfoque

Describiendo y clasificando regiones de una simple segmentación

1. SVM fueron implementadas en LIBSVM [1] nos proporciona $P(c_r = k|r)$, la probabilidad que la etiqueta de una región c_r sea k, y en nuestra clasificación de regiones es $argmax_k P(c_r = k|r)$
2. Posición de la región esta dada por un centroide normalizado por las dimensiones de la imagen.
3. Color es descrito por un histograma de dimensión de 100 características de tono.
4. La estructura de la imagen dentro o cerca de una región por un RCF(región-based context feature) de dimensión 300.
5. Histograma de distancia pesada de SIFTs cuantificados.
6. También se agrega el color y RCF sobre la imagen completa, teniendo 802 características.

Núcleo del Enfoque

Describiendo y clasificando regiones de una simple segmentación

1. ¿Existe una combinación de segmentación/parámetro/característica que nos de la mejor partición para todas las imágenes?
2. La mejor segmentación total tiene menor exactitud en la clasificación que la peor segmentación en algunas clases.
3. Una sola segmentación es desventajoso, todas las segmentaciones son la mejor o la peor en al menos una imagen, y la mayoría de las segmentaciones son la mejor y la peor en al menos una clase de objetos,
4. Ninguna segmentación debería ser descartada ya que podría producir resultados útiles.
5. Combinar las fuerzas de todos los algoritmos de segmentación.

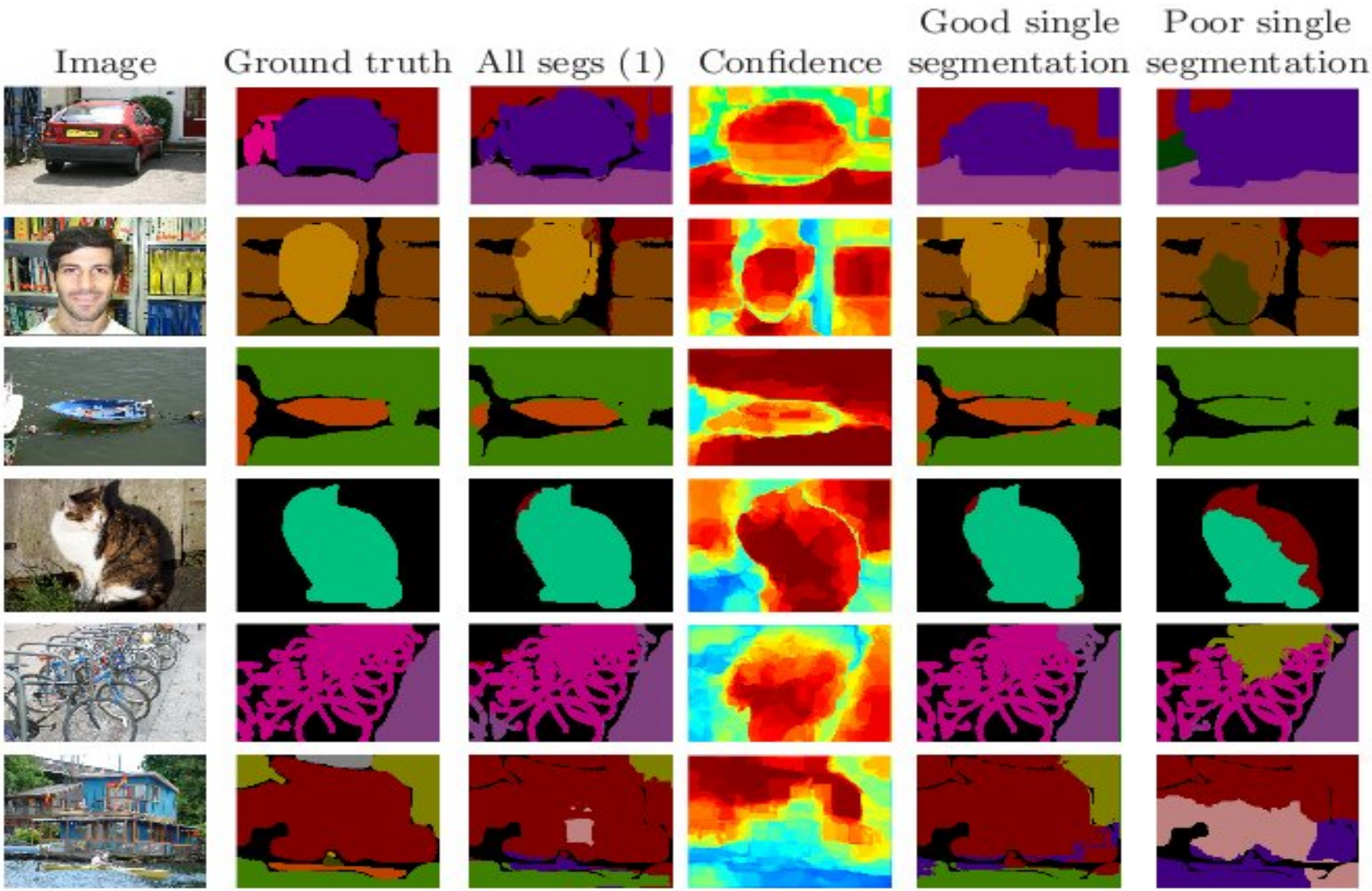

Figura: Resultados de Mapas de Objetos, cada mapa muestra el resultado más probable de cada píxel.

	class avg	**pixel avg**	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bike	flower	sign	bird	book	chair	road	cat	dog	body	boat
Shotton [14]	57.7	72.2	62	98	86	58	50	83	60	53	74	63	75	63	35	19	92	15	86	54	19	62	7
Verbeek [17]	64.0	73.5	52	87	68	73	84	94	88	73	70	68	74	89	33	19	78	34	89	46	49	54	31
Worst seg	49.6	63.3	48	80	69	51	61	87	73	71	57	47	56	34	28	15	75	16	76	28	17	40	11
Best seg	59.8	72.2	61	89	79	57	66	92	81	80	67	63	66	52	31	26	88	27	80	52	32	45	30
All segs (1)	60.3	74.3	68	92	81	58	65	95	85	81	75	65	68	53	35	23	85	16	83	48	29	48	15
All segs (2)	59.9	74.2	68	92	81	57	63	95	82	81	76	65	67	54	34	23	84	16	83	47	30	46	14

Figura: Resultados de Píxeles exactos, MS tiene mejor desempeño que Shotton et al. [5] y comparable con Verbeek [12]

.

Núcleo del Enfoque

Integrando múltiples segmentaciones

2 principios

- Píxeles los cuales son agrupados por todas las segmentaciones deberían ser clasificados de manera consistente.
- Las regiones originales proporcionan un mejor soporte para la extracción de las características que los IofRs.

1. Unidad básica son las intersecciones de regiones (IofRs), píxeles que pertenecen a la misma región en todas las segmentaciones

Núcleo del Enfoque

Integrando múltiples segmentaciones

1. Clasificar cada IofR combinando la información de todos los segmentos individuales, dado i sea un IofR, y rsi la región cual contiene i en la segmentación s. Dado c_i sea la etiqueta de la clase de i, k una etiqueta de clase especifica, I los datos de la imagen, entonces definimos el método de integración como

$$P(c_i = k|I) \propto \sum_s P(c_i^s = k|r_i^s, I)$$

2. Este promedio sobres las regiones individuales aumenta la confianza para marginalizar las regiones que contiene i, asumiendo que son igualmente probables, la clase asignada a un IofR es $argmax_k$ $P(c_i = k|I)$

Extensiones

Determinando la confiabilidad de la clasificación de una región

1. Nuestro enfoque asume que todas las segmentaciones deberían tener un voto equivalente.
2. Se asume también que la confiabilidad de la predicción de una región corresponde a el número de objetos traslapados
3. Entrenar un clasificador para predecir la homogeneidad de una región con respecto a la clase.
4. Si consideramos la homogeneidad como medida de probabilidad de una región $P(r_i^s|I)$ entonces

$$P(c_i = k|I) \propto \sum_s P(r_i^s|I)P(c_i^s = k|r_i^s, I)$$

Extensiones

Determinando la confiabilidad de la clasificación de una región

1. El clasificador usado para determinar $P(r_i^s|I)$ es un conjunto de arboles de decisión boosted, entrenados usando Adaboost,
2. Se usan 20 arboles con 16 nodos hoja para evitar sobreajuste

Características

- Posición promedio normalizada (2D)
- Histograma de color (100D)
- RCF(300D)
- Tamaño de la región dividida por el tamaño de la imagen (1D),
- Número de regiones traslapadas (1D).

3. El gasto de calcular la homogeneidad para una región es cuestionable. Brooks [6] 58.4 %, 50.1 %

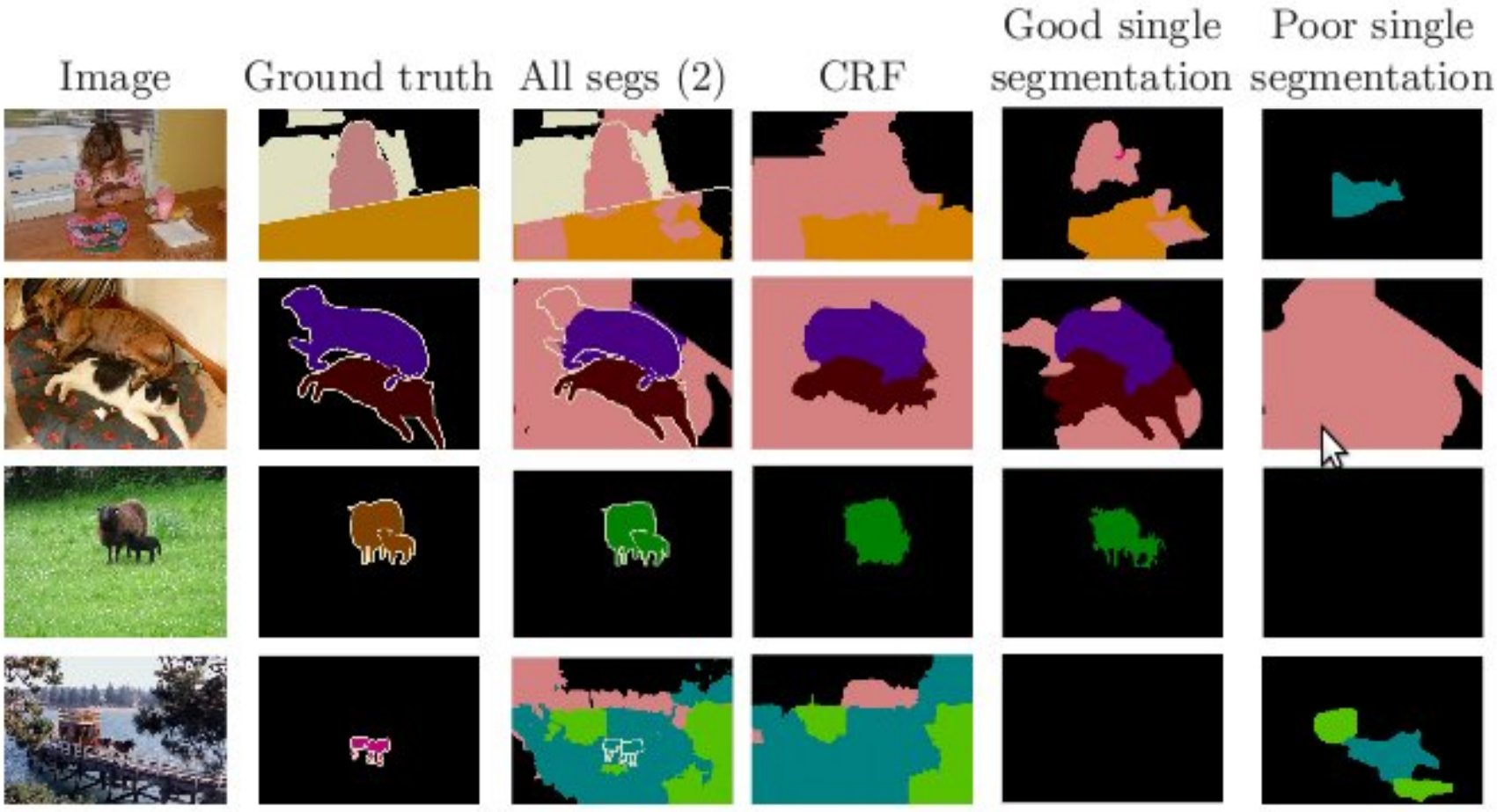

Figura: Resultados de Mapas de Objetos, cada mapa muestra el resultado más probable de cada píxel.

Extensiones

Incorporando información contextual

1. Extensión para usar información espacial explicita
2. A través (RCF), redefinimos la etiqueta de una imagen como una minimización de energía, considerando los potenciales de una simple y par de adyacentes IofRs.

$$E(C) = \sum_{i} E(c_i) + \sum_{i,j} E(c_i, c_j)$$

3. Donde C es la etiqueta de la imagen entera, *i y j* son vecinos IofRs, los potenciales unarios son definidos como $E(c_i = -logP(c_i|I))$ para penalizar incertidumbre usando confianza.

Extensiones

Incorporando información contextual

1. Los potenciales binarios, penalizan discontinuidad entre etiquetas adyacentes como sugiere [9]

$$E(c_i, c_j) = \begin{cases} 0 & \text{if } c_i = c_j, \\ \beta\left(\log p_{ij} - \log\left(1 - p_{ij}\right)\right) & \text{otherwise.} \end{cases}$$

2. Asegura que $E(c_i, c_j) > 0$ y usamos podado de grafos con expansión alfa para minimizar la energía [7].

Extensiones

Incorporando información contextual

1. P_{ij} refleja la probabilidad que las regiones padres de un IofRs pertenezca al mismo objeto en cada segmentación.

$$p_{ij} \propto \sum_s p_{ij}^s, \quad p_{ij}^s = \begin{cases} 1 & \text{if } r_i^s = r_j^s \\ P\left(c_i^s = c_j^s | r_i^s, r_j^s, I\right) & \text{otherwise} \end{cases}$$

2. Clasificadores para $P(c_i^s = c_j^s | r_i^s, r_j^s, I)$ son aprendidos con Adaboost Logístico, con las siguientes características.

Extensiones

Incorporando información contextual

1. La unión de 2 regiones es descrita usando posición promedio normalizada, RCF, histograma de color.
2. Para comparar 2 regiones, dividimos las regiones más pequeña entre el tamaño de la región más grande.
3. La divergencia KL simétrica entre las regiones individuales RCFs escaladas entre 0 y 1.
4. La divergencia KL entre los 2 histogramas de color y la diferencia normalizada de la posición de las regiones.

Extensiones

Incorporando datos entrenados débilmente etiquetados

1. Los datos entrenados han sido etiquetados completamente, pero ¿cuando el conjunto de imágenes es enorme?.
2. Se Usan datos débilmente etiquetados para incrementar el conjunto de entrenamiento. 422 - 822
3. Las etiquetas débiles toman 2 formas, bounding boxes y etiquetas de objetos a nivel de imagen el cual no contiene información de localización.
4. Se asume que las etiquetas débiles son mascaras de objetos ruidosas, si múltiples bounding boxes o etiquetas existen para un píxel, entonces son consideradas correctas.
5. La exactitud incrementa de 3-5 % para la simple y la múltiple segmentación

Conclusiones

1. Se presentó un método para reconocimiento y segmentación de objetos
2. El método consta de múltiples bottom-up segmentaciones para apoyar el reconocimiento de objetos top-down.
3. Agregando características a través de múltiples segmentaciones se incrementa la robustez.
4. Agregaron extensiones y se estudió si eran benéficas
5. Modelando la fiabilidad de la región se mejoro el desempeño.

Conclusiones II

1. Incrementando el conjunto de entrenamiento también mejora los resultados significativamente.
2. Incorporando información espacial no valió mucho la pena dada la información espacial implícita
3. Implemento clasificación de imagen y detección de objetos como un paso previo para la segmentación de objetos.
4. La importancia de la robustez de los algoritmos
5. La importancia de examinar si extensiones sobres los algoritmos recompensan su complejidad con un desempeño mejorado

Conclusiones III

1. La medición sobre las extensiones no queda bastante clara
2. Vale la pena agregar extensiones no triviales para mejorar un porcentaje no muy significativo.
3. En algunas extensiones no mencionan como se realizó la clasificación
4. Se enfocaron más en demostrar que la multiple segmentación mejora la simple segmentacón.
5. No me queda claro la idea de como la detección de objetos puede prepararse para realizar la segmentación

Unnikrishnan, R., Pantofaru, C., Hebert, M.: Toward objective evaluation of image segmentation algorithms. PAMI 29 (2007)

Malisiewicz, T. Efros, A.A.: Improving spatial support for objects via multiple segmentations. In: BMVC (2007)

A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In Proceedings of the 8th European Conference on Computer Vision, volume II, pages 71–84, 2004.

Bryan C. Russell, Alexei A. Efros, Josef Sivic, William T. Freeman, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In Proc. CVPR, 2006.

Jamie Shotton, John Winn, Carsten Rother, and Antonio Criminisi. Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation. In ECCV, 2006.

Ladicky, L., Kohli, P., Torr, P.: Oxford Brookes entry, PASCAL VOC 2007 Segmentation Challenge (2007).

Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. PAMI 23, 1222–1239 (2001)

Tomasz Malisiewicz and Alexei A. Efros, Improving Spatial Support for Objects via Multiple Segmentations, BMVC 2007

Hoiem, D., Efros, A., Hebert, M.: Recovering surface layout from an image. IJCV 75 (2007)

J. Shi and J. Malik. Normalized cuts and image segmentation. IEEE Trans. PAMI, 22(8):888–905, August 2000.

D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. IEEE Trans. Patt. Anal. Mach. Intell., 24(5):603–619, 2002.

Verbeek, J., Triggs, B.: Region classification with markov field aspect models. In: CVPR (2007)

Gracias por su Atención

¿Dudas, comentarios, sugerencias?

1. Textons refer to fundamental micro-structures in natural images, consists of a number of image bases with deformable spatial configurations
2. RCF: características de contexto basadas en regiones, es un histograma (cuantificado) de descriptores locales cercanos a una región