Matching Words and Pictures

Dan Harvey & Sean Moran

27th Feburary 2009

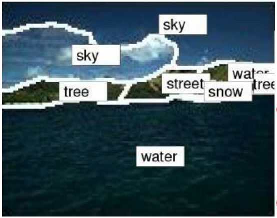
Outline

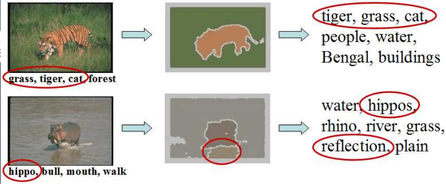
Motivation

- Images are a core part of the modern world.
- Recent explosion in number of images being captured and shared:
  - Number of images on internet estimated to be in excess of $1.5x10^{10}$.
  - Global annual sales: $1x10^{8}$ digital cameras and $3x10^{8}$ camera phones.
- Newspaper archives, picture libraries, etc maintain huge private collections.
- Great interest in how we can analyse images to ensure ease of search and browsing.
- Automatic matching of words to pictures is a potentially huge growth area.

Matching words to pictures

- Interesting application of multi-modal data mining.
- Two main types:
  - Auto-annotation: predict annotation of images using all information present.
  - Correspondence: associate particular words with particular image substructures.
- Focus on auto-annotation in this presentation.





Automatic Image Annotation

- Two main philosophies [9],[10]:
  - Block-based: Segment images and apply statistical models to those segmented regions. Most common approach in the literature e.g.:
    - CRM model of Lavrenko et al. [11]
    - Machine translation model of Duygulu et al. [12]
  - Global-feature based: Bypass segmentation stage and model global image statistics directly e.g.:
    - Robust non-parametric model of Yavlinksy et al. [10].
- Core issues for any approach:
  1. Representation: How to represent image features?
  2. Learning: How to form the classifier from training data?
  3. Annotation: How to use the classifier for novel image annotation?

Statistical Machinery

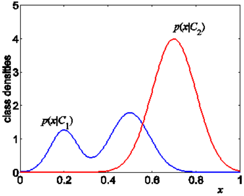
$$p(tiger \mid image)$$

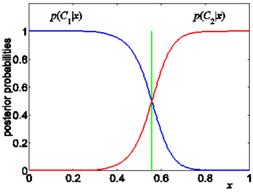
vs.

$$p(no\ tiger \mid image)$$

Bayes rule:

$$\underbrace{\frac{p(tiger \mid image)}{p(no\ tiger \mid image)}}_{\text{posterior ratio}} = \underbrace{\frac{p(image \mid tiger)}{p(image \mid no\ tiger)}}_{\text{likelihood ratio}} \cdot \underbrace{\frac{p(tiger)}{p(no\ tiger)}}_{\text{prior ratio}}$$





Key Challenges

Semantic Gap

"Lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation" [4]

Nature of Images

"Image understanding is one of the most complex challenges in AI." [5]

Scale

Occlusion

Magritte, 1957

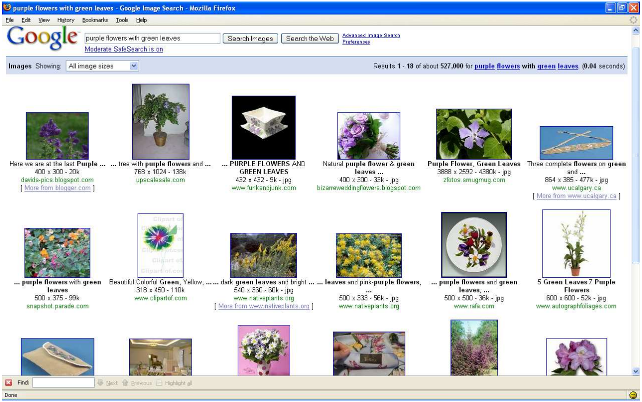
Auto-Annotation Applications

- Three core applications:
  1. Content Based Image Retrieval (CBIR) - retrieve images based on actual image content.
  2. Browsing Support - provide user with an easy way of browsing similar items.
  3. Auto-illustration - suggest pictures that might go well with surrounding text.

- Large disparity between user needs and what technology supplies e.g.:
  - Query: *"Feature is about deodorant so person should look active, not sweaty but happy, carefree - nothing too posed or set up - nice and natural looking."*[6]
  - Response: *"I'm Sorry, Dave. I'm Afraid I Can't Do That"* :-)

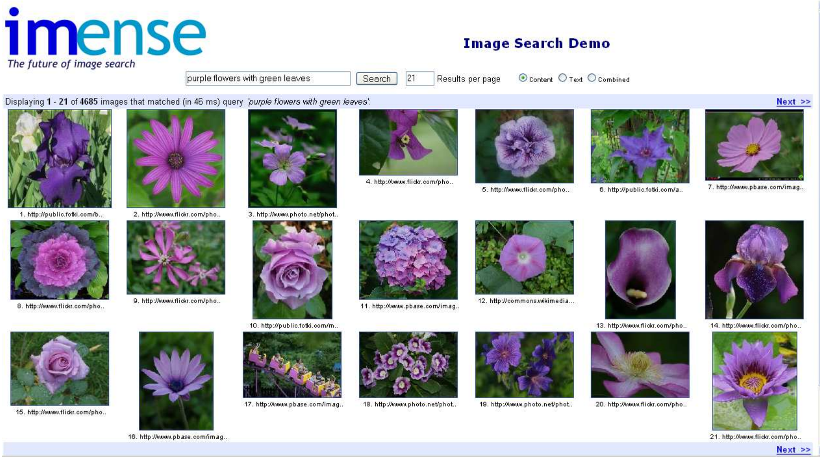
Google Image Search

- Google uses filenames, surrounding text and ignores contents of the images hence the poor retrieval results e.g. "purple flowers with green leaves":



Imense.com PictureSearch

- The Imense CBIR (www.imense.com) engine takes into account the actual image content:

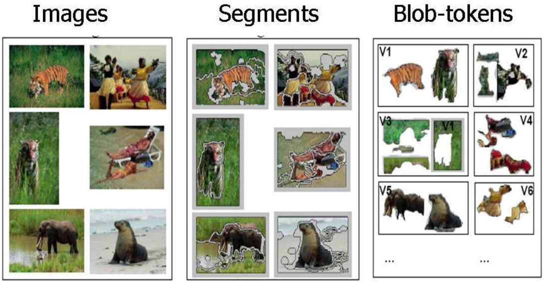


Outline

Preprocessing: How to represent an image?

- Native dimension of images is too high.
  - Resolution 481x321 = 154,401 pixels.
  - Each pixel has 3 attributes R, G, B with 255 possible values.
  - That's half a million attributes!
- Find different regions by segmentation.
- Extract features to describe each region.
- Region and features together are known as a blob.

Segmentation into regions



- Normalised Cuts (Shi and Malik, 2000)
  - Complete graph with pixels as vertices.
  - Weights on edges based feature similarity. e.g. Intensity, Colour value.
  - Recursively apply minimum cut, normalised by the number of edges cut.
- Segmentation occasionally produces small unstable regions.
- Pick 8 largest regions for feature extraction.

Geometric Features

Size

Proportion of region area to image area.

Position

Normalised coordinates of centre of mass.

Shape

1. Ratio of region area to perimeter squared.
2. Moment of inertia about centre of mass.
3. Ratio of region area to convex hull.

Other Features

Colour

Represented by average and standard deviation of :-

1. (R, G, B) Representing physical colour.
2. (L, a, b) Lightness, colour-opponent space. Representing human vision.
3. Chromaticity coordinates. Measures the quality of a colour.

$$r = \frac{R}{R+G+B} \qquad g = \frac{G}{R+G+B} \tag{1}$$

Texture

1. 4 difference of Gaussian filters.
2. 12 oriented filters at 30 degree increments.

Not the only features but a *good selection*!

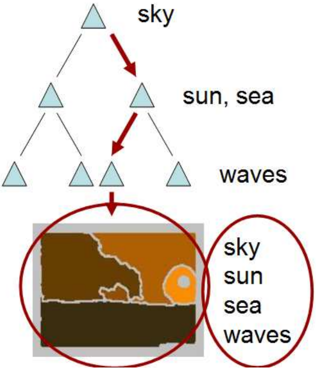
Outline

Multi-Modal Hierarchical Aspect Model

- Generative hierarchical model, combining Aspect model with a soft clustering model (Barnard & Forsyth 2001) [6][7][8]:
  - Aspect model: Models joint distribution of documents (sequence of words and image blobs) and features.
  - Soft clustering model: Maps documents into clusters.
- Images and words generated by a fixed hierarchy of nodes:
  - Leaves of the hierarchy correspond to clusters.
  - Each node has some probability of generating each word (modelled as a Multinomial distribution).
  - Each node also has some probability of generating an image segment (modelled as a Gaussian distribution).
- Images belonging to a cluster are generated by the nodes along the path from the leaf to the root.

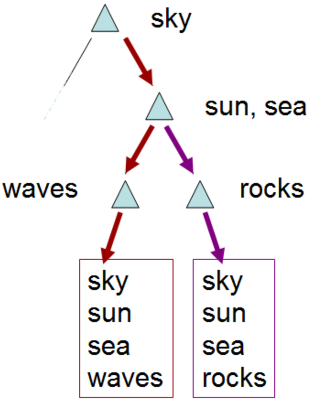
Generative nature of the Model

- Modelling data as being generated by the nodes along a path.
- For example, if the sunset image is in the 3rd cluster its words and blobs are modeled by the nodes along the indicated path:



Generative nature of the Model

- Nodes close to the root are shared by many clusters and emit items shared by a large number of data elements.
- Nodes closer to leaves are shared by few clusters and emit items specific to small number of data elements.



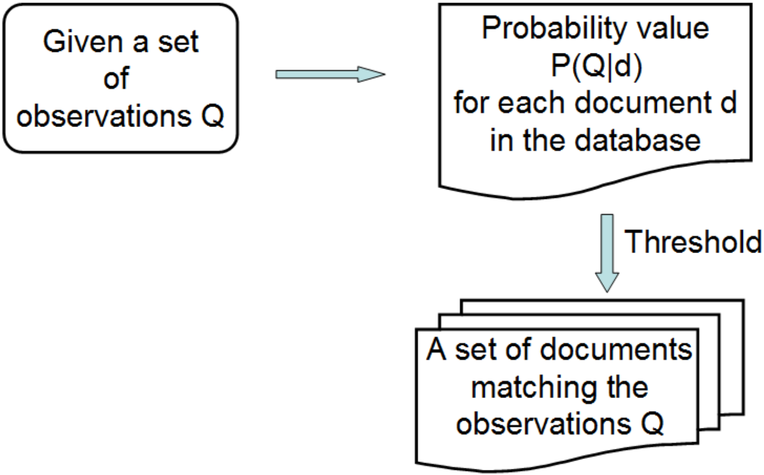
Getting technical

- A document (blobs, words) is modelled by a sum over the clusters weighted by the probability that the document is in the cluster.
- Generating a set of observations D (blobs, words) for a document d:
  - $P(D|d) = \sum_c P(c) \prod_{i \in D} \left(\sum_i P(i|l,c) P(l|c,d) \right)$
- Where:
  - c indexes clusters, i indexes items, and l indexes levels.
  - $P(i|l,c) = \text{probability of item (segment or word) in node.}$
  - $P(l|c,d) = \dfrac{\#\text{of items from node in document}}{\#\text{of document items}}$
  - $P(c,d) = \dfrac{\#\text{of document items in cluster}}{\#\text{of document items}}$
  - $P(c) = \dfrac{\sum_d P(c,d)}{\#\text{of total documents}}$

Applying the model to annotate images

- Need to calculate the probability that an image emits a proposed word, given the observed blobs, B or $P(w|B)$.
- Way to think about this conceptually:
  - Consider the probability of the items belonging to the current cluster.
  - Consider the probability of the items being generated by the nodes at various levels in the path associated to the current cluster.
  - Work the above out for all clusters.
- Mathematically:
  - $P(w|B) =$
    $$\sum_c \left(\sum_l P(w|c,l) P(l|c,B) \right) \prod_{b \in B} \left(\sum_l P(b|l,c) P(l|c) \right) P(c)$$

Applying the model to search images



Applying the model to search images

- Need to calculate the probability that a document generates a Query or $P(Q|d)$:

    - $$P(Q|d) = \sum_{c} \left(\prod_{q \in Q} \left(\sum_{l} P(q|l,c)P(l|c,d) \right) P(c) \right)$$

- Documents with a high score for $P(Q|d)$ are returned to the user.
- Soft query system: all words do not have to occur in each image returned.

Applying the model to browse images

- Browsing from coarse to fine granularity using tree structure:
- Ocean
    - Dolphins
    - Whales
    - Corals
    - and so on....

- Ocean
    - Dolphins
        - Tale
        - Head
        - and so on....

Outline

How to evaluate annotation performance?

- Compare to annotated images, not used for training.
- Show non-trival learning. (sky, water) common (tiger) uncommon.
- Performance relative to empirical word frequency.

Quality of words predicted

-ve worse, +ve better.

$$E_{KL}^{model} = \frac{1}{K} \sum_{w \in observed} log \frac{p(w)}{p(w|B)}$$

$$E_{KL} = \frac{1}{N} \sum_{data} (E_{KL}^{empirical} - E_{KL}^{model})$$

Performance Measurements

Word prediction measure

Loss function, 0 all or nothing, 1 correct, -1 compliment.

$$E_{NS}^{model} = \frac{r}{n} - \frac{w}{N - n}$$

$$E_{NS} = E_{NS}^{model} - E_{NS}^{empirical}$$

Simpler word prediction measure

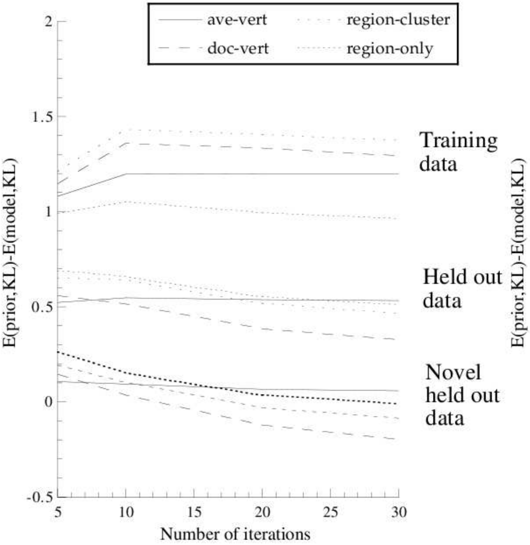
0 bad, 1 good.

$$E_{PR}^{model} = \frac{r}{n}$$

Experiments

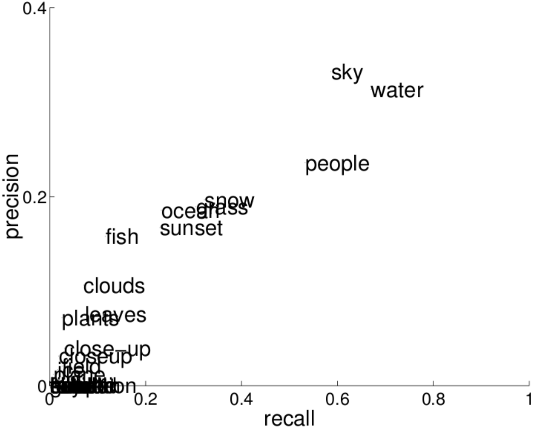
- Data set
  - Corel image data set, 160 CD's each on a specific topic. e.g. Aircrafts
  - Sample of 80 CD's, 75% training set, 25% test set
  - Remaining images were a more difficult held out set.
- Exclude words with a frequency less than 20, vocabulary of 155 words.
- 10 iterations of the training algorithm.

Experiments

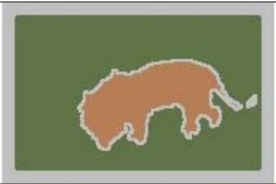


Clustering performance

Precision - Recall: Comparison



Results

		Keywords GRASS TIGER CAT FOREST Predicted Words (rank order) tiger cat grass people water bengal buildings ocean forest reef
		Keywords FLOWER coralberry LEAVES PLANT Predicted Words (rank order) fish reef church wall people water landscape coral sand trees

Results

"Methods which use image clustering are very reliant on having images which are close to the training data."

- Test set performed better than the novel held out set.
- Performs well clustering simular images.
- Less frequent and unseen blobs have lower performance.

Conclusions

- Matching words to pictures is a form of multi-modal data mining.
- Pre-process by segmenting images into feature vectors.
- Predict words for novel images by calculating $P(word|image)$.
- Multi-Modal Hierarchical Aspect Model could annotate, search and browse image collections.
- Model showed good performance on test set. Less well on the held out set.
- Exciting progress has been made, but much more work to be done!

References

1 J. Jeon, V. Lavrenko and R. Manmatha. (2003) Automatic Image Annotation and Retrieval using Cross-Media Relevance Models. *In Proceedings of the 26th Intl. ACM SIGIR Conf.*, pages 119.126, 2003.

2 K. Barnard and D. Forsyth. (2003) Learning the Semantics of Words and Pictures. *Proc. International Conference on Computer Vision.*, pp. II:408-415, 2001.

3 T. Hofmann. Learning and representing topic. A hierarchical mixture model for word occurrence in document databases. *Proc. Workshop on learning from text and the web.*, CMU, 1998.

4 A.W.M., Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain: Content based image retrieval at the end of the early years. *IEEE Trans. PAMI*, 22 (2000) 1349-1380.

References

5 M. Sonaka, V. Hlavac, R. Boyle. Image Processing, Analysis, and Machine Vision. *Brooks/Cole Publishing*, Pacific Grove, CA, 2nd Edition, 1999.

6 K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:11071135, 2003.

7 K. Barnard, P. Duygulu and D. A. Forsyth. Clustering art. *In IEEE Conf. on Computer Vision and Pattern Recognition*, II: 434-441, 2001.

8 K. Barnard and D. A. Forsyth. Learning the semantics of words and pictures. *In Int. Conf. on Computer Vision*, pages 408-15, 2001.

References

9 X. Qi and Y. Han. Incorporating multiple SVMs for automatic image annotation, *Pattern Recognition*, vol. 40, pp. 728-741, 2007.

10 A. Yavlinsky, E. Schofield, and S. Ruger. Automated image annotation using global features and robust nonparametric density estimation, *Int'l Conference on Image and Video Retrieval*, Singapore, 2005.

11 V. Lavrenko, R. Manmatha, and J. Jeon. A model for learning the semantics of pictures. In Proceedings of the 16th Conference on Advances in Neural Information Processing Systems NIPS, 2003.

12 P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. *In Seventh European Conference on Computer Vision*, volume 4, pages 97-112, 2002.