

Fast Image Auto-annotation with Visual Vector Approximation Clusters

Hervé Glotin, Sabrina Tollari

LSIS CNRS – INCOD team - 83957 La Garde cedex, France

1 Approximating Visual Clusters

We present our model "DIMATEX", a fast image auto-annotation system. Trying to tackle with the high-dimensional problem, fuzzy segmented and fuzzy labelled visual object, DIMATEX splits each visual feature in two classes : low (0) or high (1) value. See above 2 visual vector approximation clusters in 13 visual features (3 for shapes, 6 LAB colors, 4 textures):



cluster 0111010110000

cluster 1111111111110

Thus, for each image J , there is a one to one relation between each segmented region b_i and its visual cluster c_k , so we have for each word w labelling J :

$$P(w, c_k | J) = P(w, b_i | J). \quad (1)$$

2 Training Joint Visuo-Textual Distributions

Assuming that for each image J_j of training set T , $P(J_j | T) = \frac{1}{|T|}$,

$$\begin{aligned} P(w, c_k | T) &= \frac{P(w, b_i | T)}{|T|} \\ &= \sum_j P(w, b_i | J_j, T) P(J_j | T) \\ &= \frac{1}{|T|} \sum_j P(w, b_i | J_j, T) \\ &= \frac{1}{|T|} \sum_j P(w | J_j, b_i, T) P(b_i | J_j, T). \end{aligned} \quad (2)$$

We set $P(w | J_j, b_i, T) = 1$ if w annotates J_j , else 0. And we estimate for each of the m regions b_i of image J_j of T :

$$P(b_i | J_j, T) = \frac{\text{area}(b_i)}{\sum_{l=1}^m \text{area}(b_l)}, \quad (3)$$

where $\text{area}(b_i)$ is the number of pixels contained in b_i .

3 Image Auto-annotation

Using previous joint distributions, one can estimate the most accurate words for annotating an image I , by simply picking a desired number of words that have the highest probability $P(w | I, T)$ that word w annotates I . As image regions $\{b_1, \dots, b_m\}$ form a partition of I , we estimate that:

$$P(w | I, T) = \sum_{i=1}^m P(w | b_i, I, T) P(b_i | I, T), \quad (4)$$

Let be c_k the cluster of region b_i , then from(1):

$$P(w | b_i, I, T) = P(w | c_k, I, T) \simeq \frac{P(w, c_k | T)}{P(c_k | T)}, \quad (5)$$

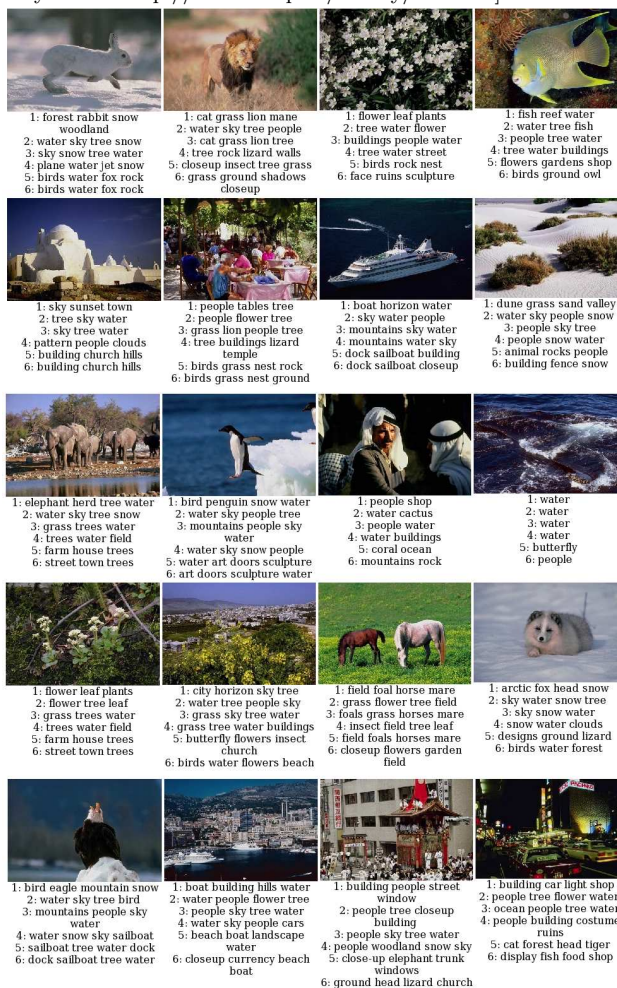
where $P(w, c_k | T)$ is given from (2), and $P(c_k | T)$ is simply estimated from the training data.

REFERENCES:

[Monay2003] F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. In Proc. of ACM Inter. Conf. on Multimedia (ACMMM2003), 2003
 [Barnard2003] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei and M. I. Jordan. Matching words and pictures. In Journal of Machine Learning Research, vol. 3, 2003

4 Results

Experiments are done on COREL where a maximum of 5 'manual' words are globally labelling each image. DIMATEX is runing segmenting each image in 10 regions with Normalized Cuts algorithm [Barnard2003]. We give below images with 1: COREL manual annotation, 2: DIMATEX Auto-Annotations, 3: PLSAWORDS, 4: PLSAWORDSFEATURES, 5: DIRECT, and 6: LSA. (3,..,6) are from [Monay2003 & <http://www.idiap.ch/monay/acmm04>]:



Gain over priors of Normalized Score (Sensi.+Specif.-1), for different auto-annotation models. Prior model uses only word frequency. DIMATEX results are given with less or more than 4 emitted words. LSA and PLSA are from [Monay2003] and Hierar. from [Barnard2003], trained and tested on the same database than DIMATEX:

