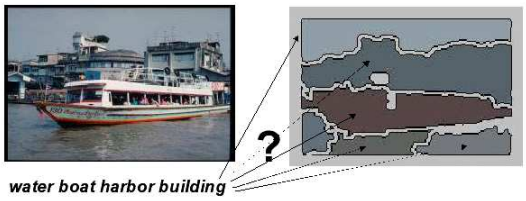
LDA versus MMD Approximation on Mislabeled Images for Keyword Dependant Selection of Visual Features and their Heterogeneity

Sabrina Tollari and Hervé Glotin
LSIS UMR CNRS – INCOD team - 83957 La Garde cedex, France. {tollari,glotin@univ-tln.fr}

1 PCA seeks for representation but LDA for discrimination

Although PCA finds components that are useful for representing data, these components are not said to be useful for discriminating between data in different classes. Considering data for 'O' and data for 'Q', PCA might ignore the tail that distinguishes an 'O' from a 'Q'.

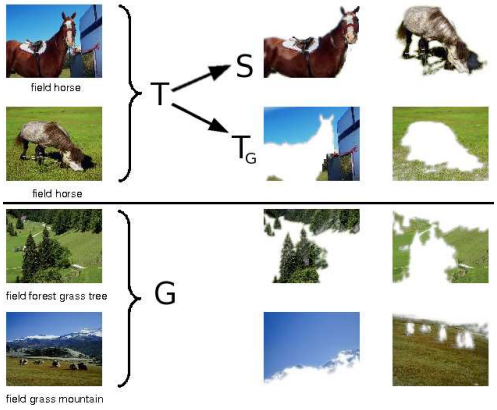
2 How to run LDA on mislabeled images ?



Due to automatic segmentation errors or abstract words, there is no one to one relationship between word and image segment (blob) in large images databases (COREL, web images...).

3 Approximation of LDA

Let be:
- S theoretical values set of one feature x for all the blobs that are *exactly* representing word w_k,
- T features set *of all blobs included in all images labeled* by w_k,
- $T = T_G \cup S$, with $T_G \cap S = \emptyset$ (we have $c_{T_G} \neq 0$; we note for set E, c_E cardinal, μ_E average of x_i values of $x \in E$, v_E variance),
- G set containing all values of x from all blobs contained in images that are not labeled by w_k,
- B_{DE} (resp. W_{DE}) Between variance (resp. Within variance) between any sets D and E.



LDA calulates for each x the discriminant power:

$$F(x;w_k) = \frac{1}{1+V(x;w_k)} \text{ where } V(x;w_k) = \frac{W_{SG}}{B_{SG}}. \quad (1)$$

We show that, if $\mu_{T_G} = \mu_G$ *and* $v_{T_G} = v_G$ (simple assumption of context independency provided by any large enougth image database):

$$\hat{V}(x;w_k) = \frac{W_{TG}}{B_{TG}} = A(w_k).V(x;w_k) + B(w_k).\Big(1 - C(x;w_k)\Big) \quad (2)$$

where $A > 0$ and $B > 0$ are *independent of* x. Moreover we show that for discriminant x $C(x;w_k) << 1$. Thus $\hat{V}(x;w_k)$ is a linear function of $V(x;w_k)$, and order of $\hat{V}$ and V values are the same. So we estimate the most discriminant features by ranking $\hat{V}$.

4 Approximation of Maximum Marginal Diversity

LDA makes the assumptions that class densities are gaussian, that are unrealistic for most problems involving real data. The best feature set characterizing word class w_i should contain those feature with large marginal diversities:

$$\hat{MD}(v_j;w_i) = \sum p(v_j|w_i) log \frac{p(v_j|w_i)}{p(v_j|\neg w_i)}. \quad (3)$$

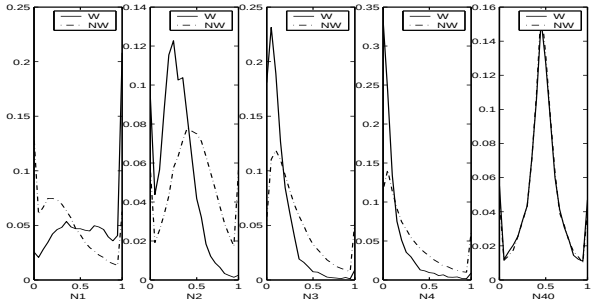


Fig. 1: Conditional likelihoods $p(v_j|w_i)$ and $p(v_j|\neg w_i)$ of 5 features for WORD (W) versus NONWORD (NW) approximated classes for keyword *SNOW*. Features are sorted from the best discriminative (N1) to the worst one (N40) (estimated by ALDA): N1 ('B' of RGB), N2 ('B' of LAB), N3 ('std A' of LAB), N4 ('std G' of RGS) and N40 ('3rd sigma texture'). We see likelihood differences for discriminant features between W versus NW classes, and overlapping for N40.

5 Adaptive Features selection

We use usual (U) features (shapes, colors, textures) and heterogenity (H) features. The heterogenity feature value of an image for a given feature x is the entropy of the values of x for all blobs of the image.

We choose the N most discriminating features which cumulate τ% of the total sum of the discriminant powers $\hat{DP}$ (= $\hat{F}$ or $\hat{MD}$) (sort by descending order):

$$\sum_{j=1}^{N} \hat{DP}(v_j;w_i) = \tau \sum_{j=1}^{\delta} \hat{DP}(v_j;w_i). \quad (4)$$

According to DEV set HAC classification results for each word, we test the fusion for each word:
- Early: concatenation of the best features
- Late: selection of best classification

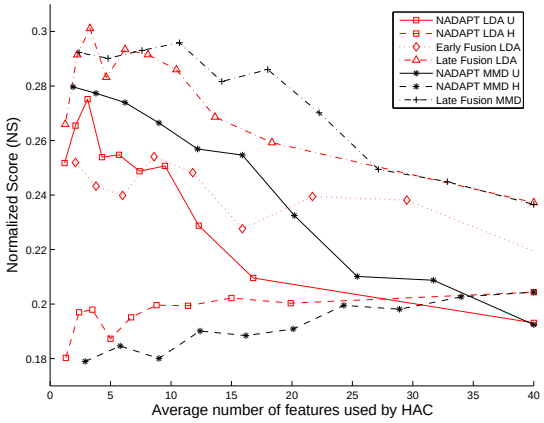


Fig. 2: Averaged NS=sensi+specif-1, over 52 keywords and 2500 TEST images, in function of the average number of features used. Each dot is for $\tau = 10\%$ to 100% (left to right). NADAPT MMD/LDA U both naturally converge to the reference model U on usual feature without feature selection (τ=1.00).