LDA VERSUS MMD APPROXIMATION ON MISLABELED IMAGES FOR KEYWORD DEPENDANT SELECTION OF VISUAL FEATURES AND THEIR HETEROGENEITY

Sabrina Tollari and Hervé Glotin

UMR CNRS 6168 LSIS System and Information Sciences lab, Université du Sud Toulon-Var
BP 20132 F-83957 La Garde cedex, France {tollari,glotin}@univ-tln.fr

ABSTRACT

We propose first to generate new visual features based on entropy measure (heterogeneity), and then we address the question of feature selection in the context of mislabeled images for automatic image classification. We compare two methods of word dependant feature selection on mislabeled images: Approximation of Linear Discriminant Analysis (ALDA) and Approximation of Maximum Marginal Diversity (AMMD). A Hierarchical Ascendant Classification (HAC) is trained and tested using full or reduced visual space. Experiments are conducted on 10K Corel images with 52 keywords, 40 visual features (U) and 40 new heterogeneity features (H). Compared to HAC on all U features, we measure a classification gain of 56% and in the same time a reduction of 92% of the number of features using a simple late fusion of U and H.

1. INTRODUCTION

Query by example is a typical mode of request in image retrieval system where user provides a query-image and the system searches for similar images based on a combination of low level multidimensional features of the query example. But this multidimensional search is not efficient due to the high dimensional problem [3, 4]. Moreover, traditional techniques in Content-Based Image Retrieval are limited by the semantic gap, which separate low-level information extracted from images and the semantic user request. Users are looking for images with semantics whereas current processing only deals with visual features. Therefore, in this paper we propose a word dependant feature selection on the usual feature space and its heterogeneity extension, to determine which are the most relevant visual features to discriminate a given concept.

In the next section, we present usual feature and new heterogeneity feature. In section 3, we describe two features selection methods: LDA and MMD, approximated for mislabeled data. In section 4, we describe the corpus and show feature selection and HAC results for usual, heterogeneity, early fusion and late fusion feature space. Finally, we discuss our results in the conclusion.

Thanks to K. Barnard [1] and J. Wang [2] for providing Corel data.

2. VISUAL FEATURES

Image processing are usually based on color, texture and shape features representing, rather roughly, major visual properties. Moreover, images are often segmented into regions (called 'blobs' that are in our paper automatically extracted by Normalized Cuts [5]; see Fig. 1). In this section we present the usual feature set we use, and we propose to generate from it a new one set motivated by psychovisual studies.

2.1. Major visual properties and usual features

Visual feature set are often chosen to be computable for any image region, and to be independent of any recognition hypothesis. As in [1], we use for each blob the 40 features listed below. Color is represented using the average and standard deviation of (R,G,B), (L,a,b), r=R/(R+G+B), g=G/(R+G+B). Texture is represented using the average and variance of 16 filter responses. We use 4 differences of Gaussian filters with different sigmas, and 12 oriented filters, aligned in 30 degree increments [5]. Shape is represented by the ratio of the area to the perimeter squared, the moment of inertia, and the ratio of the region area to that of its convex hull. Size is the image portion covered by the blob, and position is the coordinates of blob center of mass normalized by the image size. But as said in [1], it is not clear that these image features are canonical.

2.2. Heterogeneity of features

According to the experiments carried out in psychovision by J. Martinet [6], heterogeneity criterion applied to surfaces has more or less impact in visual descriptions of objects. The value of the heterogeneity of the visual feature v_j of the image d containing the b_p blobs is the entropy of the distribution of its probabilized values $b_{p,j}$:

$$H_j = -\sum_{b_p \in d} b_{p,j} \times \log_2(b_{p,j}). \quad (1)$$

In [6], heterogeneity is only defined on the area feature. Based on neurobiological studies [7], we propose in this paper to extend heterogeneity concept to all features. Recent

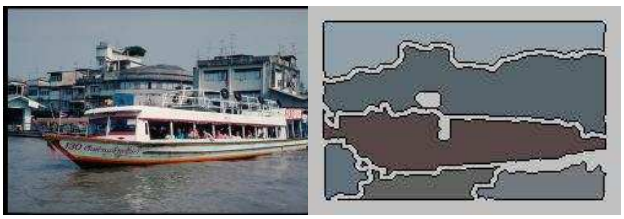

Fig. 1. Example of automatic image segmentation by Normalized Cuts algorithm [5] on Corel image labeled globally by {*WATER, BOAT, HARBOR, BUILDING*}. It's difficult to know which blob may be labeled by a word. Moreover, a bijection between blobs and words is not possible.

advances in cognitive sciences claim that human interpretation is based on a contextual visual analysis. As pointed out in [7]: "This context-dependent transformation from image to perception has profound but frequently under-appreciated implications for neurophysiological studies of visual processing". Content-based image retrieval systems should take into account this context-based neuronal bases of visual scene perception. Red color can be discriminant for 'tomato', but it is much more the heterogeneities of color features that are discriminant for 'market'. Thus we extend the visual space applying heterogeneity to all normalized usual features.

3. AUTOMATIC WORD DEPENDANT FEATURE SELECTION ON MISLABELED DATA

Due to the high dimension problem [3, 4], a good visual indexing would be made up with the visual features which have the strongest discriminating capacities. To determine which are the most relevant visual features to annotate an image with a word is a difficult problem because available (mostly mislabeled) data do not correspond with traditional statistical methods requirements. Previous works showed that simple methods like LDA[1] (Linear Discriminant Analysis) or Maximum Marginal Diversity (MMD) [9] can discriminate acoustic [10] and visual features [11], but these methods were applied on well labeled corpuses describing a univocal relation between a conceptual class and a feature. The main difficulty for applying this kind of methods on large general images corpus is that they do not have a label for each blob, but a words set for an image (Fig. 1). We make however the following assumption: *if an image database presents each concept with a rather broad contextual variety, then LDA or MMD methods can estimate the N best discriminant features of each concept.* Thus, for each word, we build a bipartition of the training set: the class WORD of images which are labeled by this word and the class NONWORD of images which are not labeled by it. Fig. 2 gives some features distributions obtained on WORD and NONWORD classes for 'snow'.

[1]Whereas PCA seeks directions that are effi cient for representation, LDA seeks directions that are effi cient for discrimination ([8] p.117).

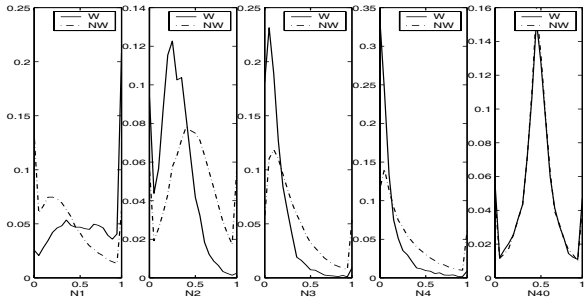

Fig. 2. Conditional likelihoods $p(v_j|w_i)$ and $p(v_j|\neg w_i)$ of 5 features for WORD (W) versus NONWORD (NW) approximated classes for keyword *SNOW*. Features are sorted from the best discriminative (N1) to the worst one (N40) (estimated by ALDA): N1 ('B' of RGB), N2 ('B' of LAB), N3 ('std A' of 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.

3.1. Approximation of Linear Discriminant Analysis

Based on our two classes WORD and NONWORD, we calculate for each word w_i and for each visual feature v_j, the between variance $\hat{B}(v_j; w_i)$ (average variance of each class) and the within variance $\hat{W}(v_j; w_i)$ (weighted average of each class variance). Finally, we calculate for each word w_i and each feature v_j the discriminant power $\hat{F}(v_j; w_i)$ defined by:

$$\hat{F}(v_j; w_i) = \frac{\hat{B}(v_j; w_i)}{\hat{B}(v_j; w_i) + \hat{W}(v_j; w_i)} \qquad (2)$$

This method, called ALDA (Approximation of LDA), has been theoretically and experimentally proved in [12, 13]. It showed that ranking errors due to this approximation are small as long as enough samples are used and considered concept is presented in various contexts.

3.2. 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 [9]. The marginal diversities $\hat{MD}(v_j; w_i)$ of feature v_j in class w_i can be defined as the Kullback-Leibler divergence between $p(v_j|w_i)$ and $p(v_j|\neg w_i)$:

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

3.3. Adaptive Features selection

To automatically determine the number of best features to discriminate each word as well as possible, we choose the N

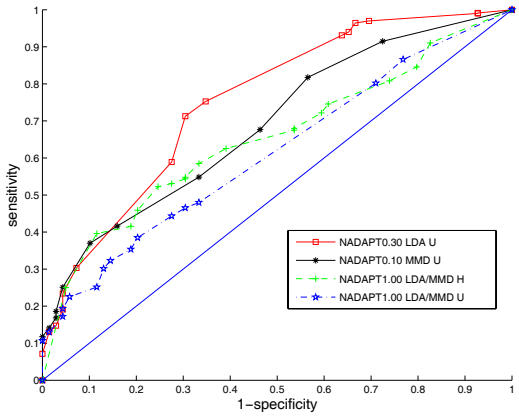

Fig. 3. ROC of the HAC image classification for word *WOMAN* applied for various methods to usual (U) and heterogeneity (H) features on DEV set. Between two points of the curve, 5% of the closest blobs are aggregated by HAC.

most discriminating ones which cumulate τ % of the total sum of the discriminant powers $\hat{DP}$ (= $\hat{F}$ or $\hat{MD}$) over all the δ features for this word (method 'NADAPTτ'). We sort $\hat{DP}$ by descending order, then the system choose N such that:

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

4. EXPERIMENTAL RESULTS

4.1. Corpus

We use the same data as in [1]. Experiments are made on Corel images database made of various 10K images, approximately 100 000 segments (called 'blobs') are preprocessed in [1] by 'Normalized Cuts' algorithm [5]. This segmenter has the occasional tendency to produce small, typically unstable regions. We keep the 10 largest regions in each image by computing, for each region, a set of 40 features described in section 2.1. Then we calculate the heterogeneity for each visual feature. In order to avoid artifact, we normalize both U and H feature vectors in 90% of their MLE Gamma distribution. Finally, each blob is represented by a vector of 80 dimensions where each component is in $[0, 1]$. Features pdf in eq. (3) are estimated by $\sqrt{256}$ bins histograms on TRAIN set. Each image of Corel is manually labeled by an average of 3.6 words from a lexicon of 250 different words. We choose to study in this article only the 52 keywords having more than 60 occurrences in our training set. The corpus is split by chance in a training set (TRAIN) of 5000 images, a development set (DEV) of 2500 images and a test set (TEST) of 2500 images.

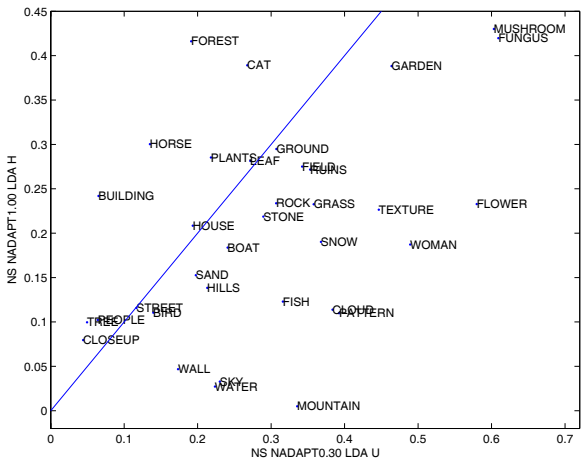

Fig. 4. Comparison of the normalized scores (NS) obtained on TEST set for NADAPT0.30 LDA U and NADAPT1.00 LDA H methods. Some words are more discriminated by heterogeneity (H) features than by usual (U) ones.

4.2. Word Clustering by HAC

To model the association between visual features for a given word, we build visual clusters by Hierarchical Ascendant Classification (HAC). For each word, we cluster by HAC the visual vectors – reduced to the N best dimensions chosen by ALDA or AMMD – of TRAIN images labeled by this word. Each visual cluster is represented by mean and std vectors. A visual cluster is an hyperrectangle in the visual multidimensional space. Finally, the system indexes an image by a word if at least 3 blobs of the image are in 1 of the visual clusters of this word ([14] for details). Each DEV and TEST image is labeled by a word set, thus we calculate the rates of sensitivity and specificity. We use the score "Normalized Score"[2] (NS= sensitivity + specificity -1) [1]. We optimize parameters (clusters sizes) on DEV set maximizing NS (Fig. 3).

4.3. Selection and fusion results on TEST

Feature selection results for 2500 TEST images and 52 keywords are shown in Fig. 5 and Tab. 1. LDA and MMD feature selections both reduce space dimension and increase score classification with U features, but not with H ones, it might be due to the lack of samples for H (only one vector by image). MMD works better than LDA with U, but not with H, nevertheless classification with H give better result than with U when all features are used. As shown in Fig. 4, some words are better discriminated with H than with U. We use this results to fusion efficiently U and H spaces. For late fusion, we learn, for each word on DEV set, which space max-

[2]NS is 1 when the system finds the n words of references and none of the other words, -1 when it only finds the words which are not references, 0 when all the words of the lexicon are found (-1$\leq$ NS $\leq$ 1).

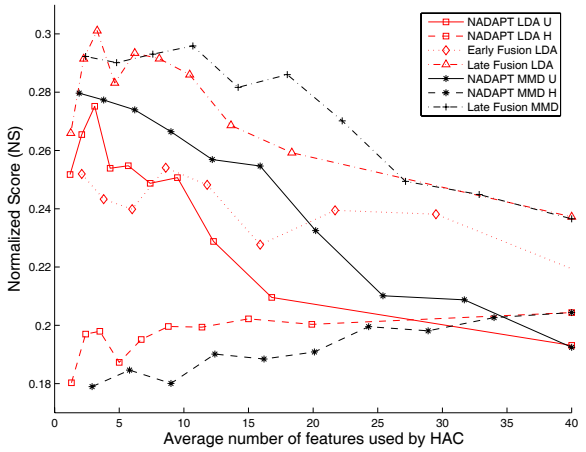

Fig. 5. Averaged NS, over 52 keywords and 2500 TEST images, in function of the average number of features used. Each dot is for $\tau = 10\%$ to 100% (left to right). NADAPT MMD/LDA U both naturally converge to the reference model U on usual feature without feature selection (τ=1.00).

imizes NS, but we keep separated LDA and MMD methods. For early fusion, for each word, we build balanced features vectors of size $Z = Z_U + Z_H$ by taking Z_U best features from NADAPT U features set and Z_H from NADAPT H one, where $Z_U/Z_H = NS_{DEV}(U)/NS_{DEV}(H)$. For example, if for one word $NS = 0.4$ by NADAPT U, and 0.2 by NADAPT H, then the late fusion vector of size 6 has the 4 best U features and the 2 best H ones.

5. DISCUSSION AND CONCLUSION

As expected NADAPT MMD U (no gaussian assumption) gives the best gain over reference: dimension is reduced by 95% with classification gain is 45%. Both late LDA and MMD fusions are more efficient than NADAPT methods because of the optimal NADAPT selection method for each words. LateLDA is better than LateMMD because NADAPT MMD H is worth than NADAPT LDA H (due to the integration eq. (1), pdf estimates are wrong). Finally, the simple and low cost LateLDA fusion is the best method reducing by 92% the visual space and enhancing by 56% the HAC. First, our paper demonstrates the efficiency of Approximated LDA or MMD features selections. Second, we demonstrate the efficiency of the late fusion of usual and heterogeneity features which are rich cues for the perceptual interpretation of ambiguous image. Stochatics fusion methods will be set for optimal fusion of these complementary visual features.

NADAPT method	τ	DIMENSION aver. N	DIMENSION reduc %	CLASSIFICATION aver. NS	CLASSIFICATION std NS	CLASSIFICATION gain %
U (ref.)	1.00	40	-	0.192	0.129	-
H	1.00	40	+0	0.204	0.116	+5
LDA U	0.30	3.1	-92	0.275	0.140	+43
MMD U	0.10	1.9	-95	0.280	0.141	+45
EarlyLDA	0.40	8.6	-79	0.254	0.120	+32
LateLDA	0.30	3.3	-92	0.301	0.133	+56
LateMMD	0.40	10.7	-73	0.296	0.131	+54

Table 1. Best results compared to reference on usual feature without selection, for 52 words and 2500 TEST images.

6. REFERENCES

[1] 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*, 2003, vol. 3, pp. 1107–1135.

[2] J. Z. Wang, J. Li, and G. Wiederhold, "Simplicity: Semantics-sensitive integrated matching for picture libraries," *IEEE Pattern Analysis and Machine Intelligence*, vol. 23, no. 9, 2001.

[3] L. Amsaleg, P. Gros, and S. Berrani, "Robust object recognition in images and the related database problems," *Multimedia Tools and Applications*, vol. 23, no. 3, pp. 221–235, 2004.

[4] K. S. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft, "When is "nearest neighbor" meaningful?," in *Proc. of ICDT, LNCS 1540*. 1999, pp. 217–235, Springer-Verlag.

[5] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[6] J. Martinet, Y. Chiaramella, and P. Mulhem, "A model for weighting image objects in home photographs," in *ACM CIKM*, 2005, pp. 760–767.

[7] Thomas D. Albright, "Why do things look as they do?: Contextual influences on visual processing," *Journal of Vision*, vol. 2, no. 10, pp. 60–60, 12 2002.

[8] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, John Wiley and Sons, Inc., 2000.

[9] Nuno Vasconcelos, "Feature selection by maximum marginal diversity: optimality and implications for visual recognition," in *Proc. of IEEE ICIP*, 2003.

[10] C. Neti, G. Potamianos, J. Luettin, I. Matthews, H. Glotin, and D. Vergyri, "Large-vocabulary audio-visual speech recognition," in *IEEE Work. Multimedia Signal Processing*, 2001.

[11] F. Zuo, P. H. N. de With, and M. van der Veen, "Multistage face recognition using adaptative feature selection and classifi - cation," in *Proc. of ACIVS, LNCS3708*. 2005, Springer.

[12] H. Glotin, S. Tollari, and P. Giraudet, "Approximation of linear discriminant analysis for word dependent visual features selection," in *Proc. of ACIVS, LNCS 3708*. 2005, Springer.

[13] H. Glotin, S. Tollari, and P. Giraudet, "Shape reasoning on mis-segmented and mis-labeled objects using approximated fi sher criterion," *Computers & Graphics*, vol. 30, no. 2, 2006.

[14] S. Tollari and H. Glotin, "Keyword dependant selection of visual features and their heterogeneity," Tech. Rep. LSIS.RR.2005.003, LSIS, 2005.