Context-based conceptual image indexing

Stéphane Ayache, Georges Quénot, Shin'Ichi Satoh

▶ To cite this version:

Stéphane Ayache, Georges Quénot, Shin'Ichi Satoh. Context-based conceptual image indexing. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2006, Toulouse. hal-00953903

HAL Id: hal-00953903
https://inria.hal.science/hal-00953903
Submitted on 3 Mar 2014

CONTEXT-BASED CONCEPTUAL IMAGE INDEXING

Stéphane Ayache and Georges Quénot[*]

CLIPS-IMAG
BP 53
38041 Grenoble Cedex 9, France

Shin'ichi Satoh

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku
Tokyo 101-8430, Japan

ABSTRACT

Automatic semantic classification of image databases is very useful for users searching and browsing, but it is at the same time a very challenging research problem as well. Local features based image classification is one of the key issues to bridge the semantic gap in order to detect concepts. This paper proposes a framework for incorporating contextual information into the concept detection process. The proposed method combines local and global classifiers with stacking, using SVM. We studied the impact of topologic and semantic contexts in concept detection performance and proposed solutions to handle the large amount of dimensions involved in classified data. We conducted experiments on TRECVID'04 subset with 48104 images and 5 concepts. We found that the use of context yields a significant improvement both for the topologic and semantic contexts.

1. INTRODUCTION

In order to retrieve images from huge digital databases, users cannot specify their needs with low-level features, but with concepts which are much more understandable. New issues in Content-Based Image Indexing (CBII) field are arising for the reduction of this semantic gap. In order to improve concept detection, many approaches take into account the context. To do so, some approaches fuse local and global descriptors [**? ?**] combined using boosted classifiers [**?**].

Other define context as spatial relationships between objects within an image [**? ? ?**] using probabilistic frameworks. However, in order to deal with a large amount of local descriptors and simplify computation, such approaches first detect points of interest and then assume them independent.

In their work, [**? ? ?**] handle semantic relationships [**?**] between concepts using Stacked Classifiers [**? ?**]. They first classify intermediate concepts, and learn their relationships in the context of a higher level concept by a second-level classifier. In TRECVID'04 experiments, [**?**] used a basis of 22 intermediate concepts. By adding the 10 TRECVID's concepts, they learned semantic context of 32 concepts with Stacking.

In this study, we propose and evaluate a new framework for incorporating contextual information into image indexing process. We assume that both semantic and local contexts can increase the accuracy of a classifier. We incorporate both local and inter-concept information in the decision process, in order to increase accuracy of classification. However, it is difficult to merge many information while managing the curse of dimensionality problem. Thus, we study an ”hybrid” approach which tries to merge contextual information with few data. We evaluate and validate our approach on five high-level features of TRECVID 2003 and 2004 Corpus.

2. MULTI-LEVEL FRAMEWORK

Our framework uses several local and global classifiers and arrange them into networks, using stacking, in order to catch contextual information. The idea is that the correlation between the input (low-level features) and the output (concepts) is too weak to be efficiently recovered by a single “flat” classifier even if the low-level features have been carefully chosen. Combining classifiers in a multi-level framework allows extraction of intermediate-level data from low-level features and other classifiers. Detection of concepts from these intermediate-level data is expected to improve the overall performance (both correctness and computation time) of the system. Figure **??** shows an overall architecture of our framework, and how classifiers are combined.

Classifiers from each level bridge a small part of the semantic gap and are expected to do it well because the correlation between their inputs and their outputs is expected to be better than the correlation between the inputs and the outputs of a single classifier that would bridge alone the whole semantic gap. Also, not only the levels are cascaded but many intermediate data are computed in parallel and the outputs of all of them are combined as inputs to the next level. This framework is able to derive high-level concepts from other intermediate concepts, but also reinforce intermediate concepts detection. Experiments of the present work will focus of the second point.

The objective of the present work is to validate our assumptions and to quantify the benefits that can be obtained from

[*] Georges Quénot is researcher at CNRS

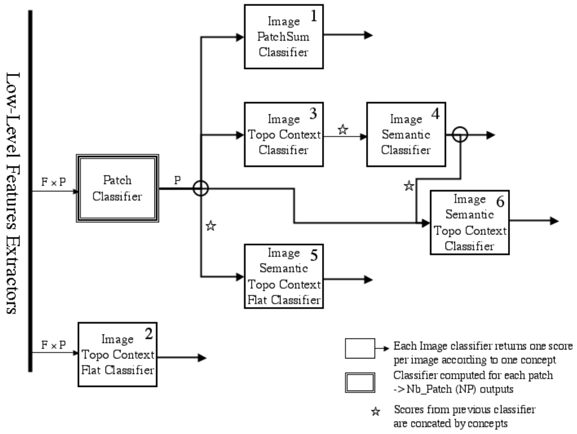
Low-Level Features Extractors
F × P
Patch Classifier
P
Image PatchSum Classifier 1
Image Topo Context Classifier 3
☆
Image Semantic Classifier 4
☆
Image Semantic Topo Context Classifier 6
☆
Image Semantic Topo Context Flat Classifier 5
F × P
Image Topo Context Flat Classifier 2
Each Image classifier returns one score per image according to one concept
Classifier computed for each patch ->Nb_Patch (NP) outputs
☆ Scores from previous classifier are concated by concepts

Fig. 1. Overall architecture and experimented networks

contextual information. We conducted several experiments with various networks of classifiers.

Figure **??** shows the multi-layer framework, the first layer (Patch Classifier, on the left) computes scores of concepts detection at patch-level. Then higher level layers assign scores for the whole images. The first layer is basically the only one that handles low-level features. The higher layers handle output of lower level layers i.e. intermediate data.

3. LOW-LEVEL FEATURES EXTRACTION

As we want to handle topologic context, we need to compute low-level features for parts of the image. In order to compute local features, many approaches have been proposed. While automatic and a priori segmented regions are too far from semantic meaning of image, we decided to split images into patches. By doing so, we should be very far from semantic, but with such granularity one patch is more likely to contain a single concept.

The low-level features extractor's process first splits image into overlapped patches. Basically, we used 260 overlapped patches of 32 × 32 pixels (in 352 × 240 images). And then, extracts 9 moments color (3 means + 6 co-variances), Gabor wavelets for texture (3 scales × 8 orientations), and coordinates of patches.

Those choices have been made for a baseline system. The main goal here is to explore the use of context for concept indexing. We want to study and evaluate various ways of doing it by combining classifiers into networks. In further work, we plan to enrich and optimize the set and characteristics of low level features, especially for video content indexing. Currently, we expect to obtain representative results from the current set of low-level features.

4. USE OF THE TOPOLOGIC CONTEXT

The idea behind the use of topologic context is that the confidence (or score) for a single patch (and for the whole image) could be computed more accurately by taking into account the confidences obtained for other patches in the image for the same concept.

We studied three network organizations to evaluate the effect of using the topologic context in concept detection at the image level. The first one is a baseline in which no context (either topologic or semantic) is used. The second one uses the topologic context in a flat (single layer) way while the third uses the topologic context in a serial (two layers) way.

In this part, we consider concepts independently one from another. Concept classifiers are trained independently from each other whatever their levels. In the following, N will be the number of concepts considered, P will be the number of patches used (260 in our experiments) and F will be the number of low-level feature vectors components (35 in our experiments).

4.1. Baseline, no context, one level (1)

In order to evaluate the patch level alone, we define an image score based on the patch confidence values. To do so, we simply compute the average of all of the patch confidence scores. This baseline is very basic it does not take into account any spatial or semantic context. We have here N classifiers, each with F inputs and 1 output. Each of them is callled P times on a given image and the P output values are averaged.

4.2. Topologic context, flat, one level (2)

The "flat" network directly computes scores at the image level from feature vectors. We have here N classifiers, each with $F \times P$ inputs and 1 output. Each of them is callled only once on a given image and the single output value is taken as the image score. This network organization is not very scalable and requires a lot of training data and training times because of the large number of inputs of the classifiers.

4.3. Topologic context, serial, two levels (3)

The "serial" network is similar to the baseline one. The difference is that the scores at the image level are computed by a second level of classifiers instead of averaging. We have here N level_1 classifiers, each with F inputs and 1 output and N level_2 classifiers, each with P inputs and 1 output. Each level_1 classifier is called P times on a given image and its P output values are passed to the corresponding level_2 classifier which is callled only once. Topologic context is taken into account by concatenating patches confidence value in a vector.

5. USE OF TOPOLOGIC AND SEMANTIC CONTEXTS

The idea behind the use of semantic context is that the confidence (or score) for a single concept could be computed more accurately by taking into account the confidences obtained for other concepts in the same image.

We studied three other network organizations to evaluate the effect of using additionnally the semantic context in concept detection at the image level. We still include outputs from the patch level, but we do so using the outputs related to all other concepts for the detection of any given concept. We are now considering concepts as related one to each other (and no longer independently one from another).

5.1. Topologic and Semantic contexts, sequential, three levels (4)

The fourth network simply take the output of the third one (topologic context, serial, two levels) and adds a third level that uses the scores computed for all concepts to re-evaluate the score of a given concept. We have additionnally here N level_3 classifiers, each with N inputs and 1 output. Each level_3 classifier is callled only once on a given image.

5.2. Topologic and Semantic contexts, parallel, two levels (5)

The fifth network is similar to the previous version except that the last two levels have been flattened and merged into a single classifier. The difference is similar to the difference between the serial and flat versions of the networks that use only the topologic context. We have here N level_1 classifiers, each with F inputs and 1 output and N level_2 classifiers, each with $N \times P$ inputs and 1 output. All level_1 classifiers are callled P times on a given image and their $N \times P$ output values are passed to the corresponding level_2 classifier which is called only once.

5.3. Topologic and Semantic contexts, parallel, three levels (6)

The previous network suffers from the same limitation as the other flattened version it is not very scalable and requires a lot of training data and training times because of the large number of inputs of the classifiers. The flattening, however, permits to use the topologic and semantic information in parallel and in a correlated way. The sequential organization, on the contrary, though making use of both information does it in a non-correlated way.

The sixth network organization try to keep both contexts correlated (though less coupled) while avoiding the curse of dimensionality problem. The $N \times P$ number of inputs is replaced by $N+P$. The architecture is a kind of hybrid between the two previous ones. It is the same as in the sequential case but P inputs are added to the classifiers f the last level. These P input comes directly from the output of the first level but for the corresponding concept only (instead as the output from all P patches times all N like in the flattened case).

6. EXPERIMENTS

We conducted several experiments in TRECVID'04 Collaborative Annotation Corpus, in order to study our framework's behavior. We used the `trec_eval` tool and TRECVID protocol, i.e. return a ranked list of 2000 top images. TRECVID-'04 Collaborative Annotation Corpus contains 48104 key frames. We split it into 50% training set and 50% test set.

We focus on 5 intermediate concepts which can be extracted as patch-level: `building`, `sky`, `greenery`, `skin_face` and `studio_setting_background`. We choose them because of their semantics relationships. `building`, `sky`, `greenery` are closer than others. Additionally, `skin_face` and `studio_setting_background` occur often together.

We used SVM classifier with RBF Kernel, because it has shown good classification results in many fields, especially in CBIR [GC04]. We use cross validation for parameter selection, using grid search tool to select the best combination of parameters C and gamma (out of 110).

In order to obtain the training set, we extracted patches from annotated regions, it is easy to get many patches by performing overlapped patches. Annotating whole images is harder as annotators must observe each one.

We collected many positive samples for patches annotation, and defined experimentally a threshold for maximum numbers of positive samples. We found that 2048 positive samples is a good compromise to obtain good accuracy with smaller training time. Also, we found that twice negative samples is a good compromise. Finally, we randomly choose negative samples. The Table **??** shows the number of positive image examples, for each concept.

Table **??** shows the relative performance and training time for the detection of five concepts and for the six considered network organizations. As expected, the flattened version requires much higher training time. For the presented times, we added the training times of each intermediate levels, it include the cross-validation time. Also, the cross-validation process can be performed in parallel [**?**], we used 11 3Ghz Pentium4 ”Hyperthreaded” processors, 1Go RAM each. The reported results are the corresponding for one single processor.

The use of topologic context improves the performance over the baseline and combined with the semantic context improves it even further. The performance of the three-level sequential classifier is poorer than the two-level serial one. This may be due to the lack of information of his final level classifier,

	build.	sky	stud.	green.	skin.	all	time
Baseline, no context, one level (1)	0.3406	0.1608	0.4087	0.6258	0.1580	0.3388	396
Topologic context, flat, one level (2)	0.1927	0.5453	0.8905	0.4623	0.3421	0.4866	836
Topologic context, serial, two levels (3)	0.3077	0.4331	0.7675	0.7207	0.4562	0.5370	418
Topo. and Semantic, sequential, three levels (4)	0.2823	0.4036	0.6498	0.7233	0.4391	0.4996	459
Topo. and Semantic, parallel, two levels (5)	0.4230	0.5606	0.9106	0.7283	0.4280	0.6101	484
Topo. and Semantic, parallel, three levels (6)	0.3381	0.4639	0.8444	0.6812	0.4424	0.5540	451
Nb positives images examples	383	1583	429	712	895		

Table 1. Comparative performance of network organizations: mean average precision (MAP) for five concepts, mean of MAPS, and corresponding training times (in minutes)

which have N (currently 5) inputs only. This may change when a much higher number of concepts are used.

For the networks which use both topologic and semantic contexts, the hybrid version has an intermediate performance between the sequential and parallel flattened versions. The two-level version has the better performance as it merge more information. However, it does not scale well with the number of concepts while the hybrid version suffers much less from this limitation and should performs better with more concepts. Also, by comparing second and fifth networks results, we can conclude that dimensionality reduction induced by our approach is really significant, in term of both accuracy and computational time.

7. CONCLUSION

We presented a framework for incorporate semantic and topologic contexts in CBII. We compared several networks and showed that both contextual information improves concepts detection. We proposed an hybrid network which is promising for further scalable experiments. Then, we found that dimensionality reduction induced by our framework, provides better accuracy in shorter computation time than a flattened classifier. This result is also very useful for further scalable experiments. Finally, we plan to evaluate how our framework can derive higher-level semantic concepts.

8. ACKNOWLEDGMENTS

This work has been spported by the ISERE CNRS ASIA-STIC project and the Video Indexing INPG BQR project.

References

[] K. Murphy, A. Torralba, D. Eaton and W. Freeman Object detection and localization using local and global features. Sicily Workshop on Object Recognition, 2005. Lecture Notes in Computer Science (submitted)

[] A Garg, S Agarwal, T.S. Huang Fusion of Global and Local Information for Object Detection. In 16th International Conference on Pattern Recognition (ICPR'02) - Volume 3, 2002.

[] J. Amores, N. Sebe, P. Radeva, T. Gevers and A. Smeulders Boosting Contextual Information in Content-Based Image Retrieval. In MIR, 2004.

[] A. Singhal, J. Luo, and W. Zhu Probabilistic spatial context models for scene content understanding. In CVPR, 2003.

[] A. Torralba, K. Murphy, and W. Freeman Contextual models for object detection using boosted random felds. In Advances in Neural Info. Proc. Systems, 2004.

[] P. Carbonetto, N. d. Freitas, and K. Barnard. A statistical model for general contextual object recognition. In ECCV, 2004.

[] G. Iyengar and H.J. Nock Discriminative model fusion for semantic concept detection and annotation in video. MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia, 2003.

[] M. Naphade. On supervision and statistical learning for semantic multimedia analysis. Journal of Visual Communication and Image Representation, 15(3):348369, 2004.

[] C.G.M. Snoek, M. Worring, J.M. Geusebroek, D.C. Koelma and F.J. Seinstra The MediaMill TRECVID 2004 Semantic Video Search Engine. InTRECVID Workshop, 2004.

[] M. Fink P. Perona Mutual boosting for contextual influence. In Advances in Neural Info. Proc. Systems, 2003.

[] D.H. Wolpert Stacked Generalization. Neural Networks, Vol. 5, pp. 241-259, Pergamon Press.

[] D. A. Lisin, M. A. Mattar, M B. BlMark C. Benfield and E. G. Learned-Miller Combining Local and Global Image Features for Object Class Recognition In CVPR, 2005.

[] P. Howarth and S. Rueger. Evaluation of Texture Features for Content-Based Image Retrieval. In P. Enser et al. (Eds.): CIVR 2004, LNCS 3115, pp.326-334, 2004.

[] P.H. Gosselin and M. Cord. A Comparison of Active Classification Methods for Content-Based Image Retrieval. Int. Workshop on Computer Vision Meets Database, CVDB, 2004.

[] C. Chang and C Lin LIBSVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm