Fine-grained recognition by sequential hypothesis rejection and foveated vision on parts

Stéphane HERBIN

ONERA – The French Aerospace Lab
F-91761 Palaiseau, France
stephane.herbin@onera.fr

Abstract. Visual foveation is the association of a spatially variant resolution sensor – the retina – and a dynamic controlled mechanism for directing the area of maximal acuity. This paper describes a processing chain able to exploit and control foveated vision for high level interpretation tasks requiring high resolution in several areas of the field of view, such as subordinate or fine-grained recognition. The process sequentially rejects wrong hypotheses by applying binary classifiers between subsets of hypotheses on local parts according to an adaptive policy maximizing the rejection capacity. The algorithm is evaluated on a problem of fine-grained car classification. Foveation is mimicked by subsampling high resolution images[1]. The underlying addressed question is the design of high-level image understanding tasks that are compliant with a given visual bandwidth, i.e. with a given budget of acquired pixels per time.

1 Sequential hypothesis rejection

The recognition task is considered as a sequential hypothesis rejection process, starting from a set of possible hypotheses or classes Ω_0 and iteratively reducing the current set of active hypotheses Ω_t by applying a sequence of tests to selected piece of data, here discriminative parts of an object. Several tests, with different rejected hypotheses, can be applied to the same data, and at a different time of the process.

The main difference with the usual way sequential testing is performed is that what is maintained during the process is not the posterior probability on the whole set of hypotheses, or a function of it (entropy, mutual information), but simply the estimated support of this distribution, i.e. the set of active hypotheses. Tests are chosen according to their *rejection capacity* as described below.

It is assumed the availability of a repertoire of tests indexed by k. A rejection test k is a function w_k with values in $\{0,1\}$ of a data or a feature X. The function is designed to be a negative indicator of a set of hypotheses Ω_k^-: if its output value is 0, all hypotheses $Y \in \Omega_k^-$ can be discarded with probability 1. This behavior can usually be achieved by setting an appropriate threshold. We define $\beta_k(Y) = \mathrm{E}_{X|Y}[w_k(X) = 0]$ the hypothesis-wide power of the rejection test for $Y \notin \Omega_k^-$ where $\mathrm{E}_{X|Y}$ is the mathematical expectation on data X when the hypothesis Y is true. The bigger value of $\beta_k(Y)$, the higher chance hypotheses in Ω_k^- are rejected by test k if Y is true.

[1] A short video illustrating the recognition process is available at http://youtu.be/51IbY3A0yC4

The average rejection capacity of test k when the current set of active hypotheses is Ω can be defined as:

$$\bar{C}_k(\Omega) = |\Omega_k^- \cap \Omega| \sum_{Y \in \Omega \backslash \Omega_k^-} \pi_Y \beta_k(Y) \tag{1}$$

where π_Y is the prior on hypothesis Y. This quantity is proportional to the number of potentially rejected hypotheses and to the average power over potentially hypotheses. With a little algebra and a uniform prior, it is easy to show that the best test is the one that maximizes $|\Omega_k^- \cap \Omega|.|\Omega \backslash \Omega_k^-|$, i.e. a test that tries to reject half of the active hypotheses.

Given a repertoire of tests $k \in \mathcal{K}$, and their associated rejection capacity function $\bar{C}_k(\Omega)$, the sequential recognition by rejection algorithm can be easily implemented following a greedy policy $k^*(t) = \text{argmax}_{k \in \mathcal{K}} \bar{C}_k(\Omega_t)$.

The testing strategy used in the experiments follows a simple adaptive scheme function of the current set of active hypotheses Ω_t. The global process maintains an accumulation counter registering the number of times each hypothesis has passed a rejection test. The tests are ranked according to 1 and applied sequentially. Hypotheses that have been considered for rejection more than τ times are removed from Ω_t and a new ranking is computed from the updated set Ω_{t+1}. The role of such an accumulation scheme is to avoid bad estimation of the rejection thresholds when learning data is scarce. A typical value of 5 was empirically proven to be optimal on the problem tested.

The process stops either when allowed time is exhausted, or when the repertoire of valid actions has been exhausted or when at most one hypothesis remains active. If more than one hypothesis remains active, the prediction is the hypothesis with the least number of votes for rejection in the accumulator.

2 Related work

Local attributes and parts for fine-grained recognition One of the key questions in fine grained visual categorization is the construction of the good features able to deal with the new discrimination vs. invariance tradeoff required for this type of problem [1–5]. A series of work propose to use basic level categorization to locate the informative parts that can be used for the subordinate classification. In [6] a strongly supervised DPM [7] is used to detect the parts. [8] describes a non adaptive sequential approach exploiting and constructing features both at the basic and subordinate category level applied to leaf recognition.[9] uses shape alignment at basic-level categorical to help position more specific features. [10] present an approach for fine-grained recognition of cars and extend their previous work by introducing geometric 3D modeling in the feature extraction process. [11] develops an original approach for the identification of high resolution informative parts from human expertise by providing to the user blurred images.

Sequential decision process for object recognition Sequential decision strategies have a long history in statistics since the early work of Wald [12] (see [13] for recent developments). In computer vision, sequential decision processes have been implemented in the form of coarse to fine strategies [14–16] or cascade-like structures [17] applied to categorical object detection rather than classification. Fewer studies have addressed the question of object classification.[18] describes a random sampling strategy optimizing the asymptotic speed of convergence of a sequential maximum likelihood test. In [19–22] a policy able to select the best detector or features to apply at a given location is learned by reinforcement learning. In artificial and robotics systems, sequential attention [23–25] and perception/action loops have been a traditional topic of investigation. They have been mostly reduced to the question of "where to look next?" [26] in order to search for a given object [27–29] or to build a representation of the environement [30]. Higher-level cognitive functions have been often reduced to an active object recognition where the main goal was to move the sensor in order to acquire a more informative viewpoint on a 3D object [31–33].

3 Experiments

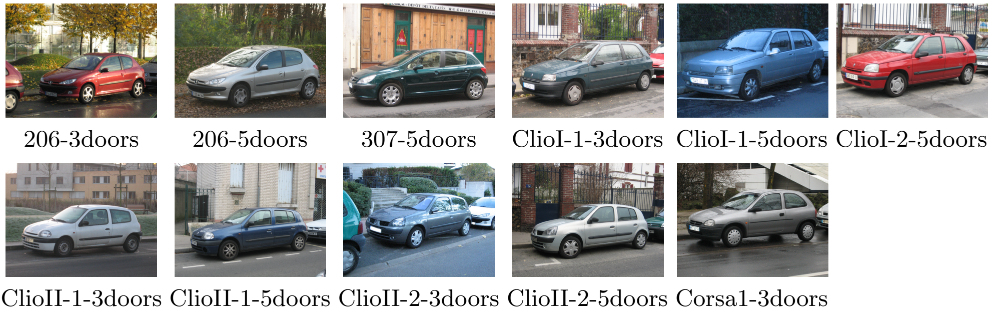

Fig. 1. Samples of the 11 classes of the image database. The high resolution images have size 1600 × 1200.

The recognition by sequential rejection approach has been applied to foveated vision for fine grained classification of cars. Foveation is simulated by observing a high resolution image with a given subsampling ratio on specific areas.

The database of images exploited contains 166 images of 11 classes of sub-compact cars taken approximately with the same orientation (Fig. 1). They were all taken with the same camera and with similar focal length. The objects have very similar body shapes, and 7 of the classes are versions of the same model (different years, different number of doors). All the images are richly semantically annotated with 15 details or subparts of various sizes (Fig. 2). This database has been previously used in [34] for a problem of hierarchical multi-label annotation.

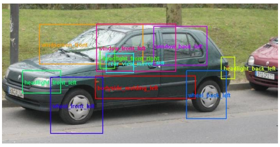

Fig. 2. Rich annotation on car.

Object parts are located relatively to a basic-level category car detector operated at low resolution. A subsampling with factor 10 of the original image, reducing the whole field of view to a 160 × 120 image gave sufficiently good results using the provided models with the available DPM distribution [35].

Once the overall object shape is detected, each part at a given scale, according to the rejection test applied, is searched inside an area defined in local coordinates from the basic-level detection bounding box. The detection of the part ("rear window mirror", "front left headlight", "front wheel" ...) is then produced by applying a detector in a sliding window approach, and estimating the part location as the response of maximal value. The width of each part, the sub-sampling factor and the size of the search area have been optimized off-line. Overall, the maximal number of pixels acquired to process all the parts and detecting the car at low resolution is compressed by a factor 16 compared to the full resolution data.

Learning occurs at two levels: categorical detection of parts and rejection tests. Since the database is small, cross validation with 15 folds ensuring that each fold contained at least one sample of each class has been used.

For the categorical detection of parts, the positive examples are taken from the annotated data, and are enhanced by generating small affine deformations of the original data in order to take into account slight viewpoint variations. Negative samples are generated by randomly sampling regions around the annotated part. The features used are a HOG like 4 × 4 concatenation of histogram of gradient orientations on local sub-windows coded in 8 directions and weighted by the gradient magnitude. The detector is learned using a linear kernel SVM.

The construction of the rejection test uses the same feature and kernel, but on a binary discrimination of two subsets of hypotheses: Ω_k^- and $\Omega_t \backslash \Omega_k^-$. The decision threshold asserting that $w_k(X) = 0$ when a hypothesis Y belongs to Ω_k^- is set by fitting a generalized logistic function to the cumulative distribution. The test powers $\beta_k(Y)$ are then computed accordingly.

The recognition performances have been evaluated on two sequential testing strategies: a random test, and the application of the greedy policy (1). The maximum value of the probability of good recognition is 74% for this 11 class fine-grained problem. A less refined problem using 5 super classes ("206", "307", "ClioI", "ClioII", "Corsa") gave 94.6% of accuracy.

Fig. 3. Performances using several policies.

	Mean accuracy (%)				
Scanpath length	20	50	100	150	200
Random policy	16.3	28.0	39.2	45.2	52.4
Greedy policy	42.2	60.4	67.5	71.2	**74.2**
Centralized policy	71.7				

To compare the approach, a centralized classification exploiting a compound feature concatenating the HOG features of all parts and using a standard one vs. all multiclass SVM with a linear kernel was computed and gave equivalent results (71.7%).

3.1 Possible improvements

The focus or search area depends on a low resolution basic-level categorical detection, which lacks spatial precision. An alternative is to use local landmarks as local coordinates or a strongly supervised global model [7]. The part descriptor can be improved and made more discriminative (conv. net features?). Finally, the greedy policy used to select the rejection tests may be replaced by a more global planning process with look ahead strategy and trying to optimize more general costs than the number of recruited rejection tests.

References

1. Yao, B., Khosla, A., Fei-Fei, L.: Combining randomization and discrimination for fine-grained image categorization. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 1577–1584
2. Farrell, R., Oza, O., Zhang, N., Morariu, V.I., Darrell, T., Davis, L.S.: Birdlets: Subordinate categorization using volumetric primitives and pose-normalized appearance. In: Computer Vision (ICCV), 2011 IEEE International Conference on, IEEE (2011) 161–168
3. Duan, K., Parikh, D., Crandall, D., Grauman, K.: Discovering localized attributes for fine-grained recognition. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 3474–3481
4. Berg, T., Belhumeur, P.N.: Poof: Part-based one-vs.-one features for fine-grained categorization, face verification, and attribute estimation. 2013 IEEE Conference on Computer Vision and Pattern Recognition **0** (2013) 955–962
5. Gao, S., Tsang, I.H., Ma, Y.: Learning category-specific dictionary and shared dictionary for fine-grained image categorization. Image Processing, IEEE Transactions on **23**(2) (Feb 2014) 623–634
6. Zhang, N., Farrell, R., Iandola, F., Darrell, T.: Deformable part descriptors for fine-grained recognition and attribute prediction. In: Computer Vision (ICCV), 2013 IEEE International Conference on, IEEE (2013) 729–736
7. Azizpour, H., Laptev, I.: Object detection using strongly-supervised deformable part models. In: Computer Vision–ECCV 2012. Springer (2012) 836–849
8. Sfar, A.R., Boujemaa, N., Geman, D.: Vantage feature frames for fine-grained categorization. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE (2013) 835–842
9. Gavves, E., Fernando, B., Snoek, C.G.M., Smeulders, A.W.M., Tuytelaars, T.: Fine-grained categorization by alignments. In: Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia (December 2013)
10. Krause, J., Stark, M., Deng, J., Fei-Fei, L.: 3d object representations for fine-grained categorization. In: 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13), Sydney, Australia (2013)
11. Deng, J., Krause, J., Fei-Fei, L.: Fine-grained crowdsourcing for fine-grained recognition. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. (June 2013) 580–587
12. Wald, A., Wolfowitz, J.: Optimum character of the sequential probability ratio test. The Annals of Mathematical Statistics **19**(3) (1948) 326–339
13. Naghshvar, M., Javidi, T.: Active sequential hypothesis testing. The Annals of Statistics **41**(6) (12 2013) 2703–2738
14. Blanchard, G., Geman, D.: Hierarchical testing designs for pattern recognition. Annals of Statistics (2005) 1155–1202
15. Fidler, S., Boben, M., Leonardis, A.: A coarse-to-fine taxonomy of constellations for fast multi-class object detection. In: Computer Vision–ECCV 2010. Springer (2010) 687–700
16. Gangaputra, S., Geman, D.: A design principle for coarse-to-fine classification. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on. Volume 2., IEEE (2006) 1877–1884
17. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR 2001. Volume 1., IEEE (2001) I:511–518
18. Herbin, S.: Active sampling strategies for multihypothesis testing. In: Energy Minimization Methods in Computer Vision and Pattern Recognition, Springer (2003) 97–112

19. Paletta, L., Fritz, G., Seifert, C.: Q-learning of sequential attention for visual object recognition from informative local descriptors. In: Proceedings of the 22nd international conference on Machine learning, ACM (2005) 649–656
20. Trapeznikov, K., Saligrama, V.: Supervised sequential classification under budget constraints. In: AISTATS. (2013) 581–589
21. Karayev, S., Fritz, M., Darrell, T.: Anytime recognition of objects and scenes. In: CVPR. (2014)
22. Dulac-Arnold, G., Denoyer, L., Thome, N., Cord, M., Gallinari, P.: Sequentially generated instance-dependent image representations for classification. In: International Conference on Learning Representations. (2014)
23. Frintrop, S., Rome, E., Christensen, H.I.: Computational visual attention systems and their cognitive foundations: A survey. TAP **7**(1) (2010)
24. Borji, A., Itti, L.: State-of-the-art in visual attention modeling. IEEE Transactions on Pattern Analysis and Machine Intelligence (2012)
25. Filipe, S., Alexandre, L.A.: From the human visual system to the computational models of visual attention: a survey. Artificial Intelligence Review (2013) 1–47
26. Roy, S.D., Chaudhury, S., Banerjee, S.: Active recognition through next view planning: a survey. Pattern Recognition **37**(3) (2004) 429 – 446
27. Sommerlade, E., Reid, I.: Information-theoretic active scene exploration. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. (June 2008) 1–7
28. Aydemir, A., Pronobis, A., Göbelbecker, M., Jensfelt, P.: Active visual object search in unknown environments using uncertain semantics. IEEE Transactions on Robotics **29**(4) (August 2013) 986–1002
29. Andreopoulos, A., Tsotsos, J.: A computational learning theory of active object recognition under uncertainty. International Journal of Computer Vision (2012) 1–48
30. Aydemir, A., Jensfelt, P.: Exploiting and modeling local 3d structure for predicting object locations. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE (2012)
31. Defretin, J., Herbin, S., Le Besnerais, G., Vayatis, N.: Adaptive planification in active 3d object recognition for many classes of objects. In: Workshop Towards Closing the Loop: Active Learning for Robotics, RSS Robotics: Science and Systems Conference. (2010)
32. Deinzer, F., Derichs, C., Niemann, H., Denzler, J.: A framework for actively selecting viewpoints in object recognition. IJPRAI **23**(4) (2009) 765–799
33. Herbin, S.: Combining geometric and probabilistic structure for active recognition of 3D objects. In: ECCV. Volume 1407 of Lecture Notes in Computer Science., Berlin, Springer Verlag (1998) 748–764
34. Tousch, A.M., Herbin, S., Audibert, J.Y.: Semantic lattices for multiple annotation of images. In: Proceedings of the 1st ACM international conference on Multimedia information retrieval. MIR '08, New York, NY, USA, ACM (2008) 342–349
35. Felzenszwalb, P.F., Girshick, R.B., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. Pattern Analysis and Machine Intelligence, IEEE Transactions on **32**(9) (2010) 1627–1645