Visual dictionaries in the Brain: Comparing HMAX and BOW

Ramakrishnan, K.; Groen, I.I.A.; Scholte, H.S.; Smeulders, A.W.M.; Ghebreab, S.

Published in:
2014 IEEE International Conference on Multimedia and Expo (ICME 2014): Chengdu, China 14-18 July 2014

DOI:
10.1109/ICME.2014.6890312

VISUAL DICTIONARIES IN THE BRAIN : COMPARING HMAX AND BOW

Kandan Ramakrishnan[⋆]*, Iris I. A. Groen*[†]*, H. Steven Scholte*[†]*, Arnold W. M. Smeulders*[⋆]*, Sennay Ghebreab*[⋆†]

[⋆]Intelligent Sensory Information Systems Group, University of Amsterdam, The Netherlands
[†] Cognitive Neuroscience Group, University of Amsterdam, The Netherlands

ABSTRACT

The human visual system is thought to use features of intermediate complexity for scene representation. How the brain computationally represents intermediate features is, however, still unclear. Here we tested and compared two widely used computational models - the biologically plausible HMAX model and Bag of Words (BoW) model from computer vision against human brain activity. These computational models use visual dictionaries, candidate features of intermediate complexity, to represent visual scenes, and the models have been proven effective in automatic object and scene recognition. We analyzed where in the brain and to what extent human fMRI responses to natural scenes can be accounted for by the HMAX and BoW representations. Voxel-wise application of a distance-based variation partitioning method reveals that HMAX explains significant brain activity in early visual regions and also in higher regions such as LO, TO while the BoW primarily explains brain acitvity in the early visual area. Notably, both HMAX and BoW explain the most brain activity in higher areas such as V4 and TO. These results suggest that visual dictionaries might provide a suitable computation for the representation of intermediate features in the brain.

Index Terms— Visual perception, fMRI, low and intermediate features, HMAX, Bag of Words, Representation Similarity Analysis

1. INTRODUCTION

The human visual system transforms low-level features in the visual input into high-level concepts such as objects and scene categories. Much is known about the computation of low-level visual features such as color and orientation [1-2]. How these low-level features are transformed into high-level object and scene percepts, however, is still the subject of research.

One possibility is that the visual system first creates intermediate represention [3] in a hierarchical fashion of the visual input, and then transforms these into full object and scene representations. In neural models of vision, such intermediate features are deemed important for scene categorization because they have a good trade off between frequency and specificity [4]. In this paper we study two models with such intermediate representations.

Currently there are two main computational models of vision, both use hierarchical visual processing for real world image categorization. The HMAX model [5], for example, is a biologically plausible model that uses features of intermediate complexity for object recognition. In computer vision, the Bag of Words model (BoW) [6] is a successful model for scene classification.

HMAX[5] performs the initial feedforward stage of object recognition in the ventral visual pathway. It extends the idea of simple and complex cells by forming a hierarchy in which alternate template matching and max pooling operations progressively build up feature selectivity and invariance to position and scale. HMAX is the widely accepted model for object recognition in the brain, however is not the preferred choice in computer vision algorithms.

The dominant form in computer vision is Bag of Words which performs very well on large TRECvid [7] and PASCAL [8] datasets, in some cases even approaching human classification performance [9]. The key idea behind this model is to quantize local SIFT features [6] into visual words, features of intermediate complexity, and then to represent an image by a histogram of visual words. BoW being so successful, it is a candidate model to test against the human brain.

Inspite of the differences, both HMAX and BoW models use the concept of visual dictionaries. In HMAX at the S2 layer, template patches are learnt from a dataset of images which are convolved with the responses from the C1 layer. These template patches are refered to as dictionary elements or visual dictionary. In the BoW, the clustering of SIFT features into words form the visual dictionary. The visual dictionary computed in these models are features of intermediate complexity, the focus of study in this paper.

In this work we hypothesize that the human visual system uses intermediate features for scene representation and we compare how HMAX and BoW models explain brain activity. We expect to find areas in the brain where activity is accounted for by HMAX or BoW representations of visual input. In particular, we expect to find this for areas beyond early visual cortex where increasingly complex information is processed. To test this, we record fMRI responses of several subjects to natural scenes, and search the fMRI volumes for voxels that are significantly explained by HMAX or BoW representations.

Finding how visual distionaries explain brain activity is challenging for two reasons. First computational and neural representations are heterogeneous and at the same time high-dimensional. Second, the representation from HMAX and BoW build on Gabor filters and SIFT features respectively, and hence these need to be dissociated in a proper manner. We address the first challenge by considering dissimilarity matrices [10], and the second is resolved by applying variation partitioning [11] on these dissimilarity matrices to compute the unique contributions of HMAX and BoW representations explaining brain responses.

2. REPRESENTATION OF VISUAL SCENES AND BRAIN RESPONSES

2.1. HMAX representation

We use the HMAX model from [5] where features are computed hierarchically in layers: an initial image layer and four subsequent layers, each built from the previous by alternating template matching and max pooling operations. In the first step, the image is downsampled and a image pyramid of 10 scales is created. Gabor filters of four orientations are convolved over the image at every possible position and scale in the next step, the S1 layer. Then in the C1 layer, these responses are maximally pooled over small regions of the image. In the next step template matches between the patch of C1 units centered every position/scale and each of d prototype patches. These d prototype patches are sampled randomly from the C1 layer forming the dictionary. In the last layer a d dimensional feature is created by maximally pooling over all scales, orientations to one of the models d patches. We denote the HMAX features of an image I_k by $\mathbf{x}_1, .., \mathbf{x}_d$.

2.2. BoW representation

The first step in the BoW model is extraction of SIFT [12] features from the visual input. Here the SIFT representation of an image I_k is denoted by $\mathbf{s}_k = \mathbf{f}_1, ..., \mathbf{f}_N$ where $\mathbf{f}_n$ is a 128 dimensional SIFT vector at N interest points in the image. We use dense sampling with 2 pixel spacing and at multiple scales.

Secondly, a dictionary of visual words [6] is learned from an independent set of scenes. We use k-means clustering to identify cluster centers $\mathbf{c}_m = \mathbf{c}_1, ..., \mathbf{c}_M$ in SIFT space, where $m = 1, ..., M$ denotes the number of clusters centers.

All SIFT features of a new image are assigned to the most similar word and the image is represented by counting the occurrences of all words. This results for image I_k in a visual word histogram $\mathbf{w}_k = h_1, ..., h_M$ where each bin h_m indicates the number of times the word $\mathbf{c}_m$ is present in the image. We use the PASCAL VOC 2007 [8] dataset to create a codebook of dimension $M = 4000$.

2.3. fMRI response representation

Participants (n=4) repeatedly viewed (9 runs) a set of 72 scenes drawn from 6 scene categories: beaches, forests, city scenes, industry, mountains and highways [13] while we recorded BOLD-MRI (GE-EPI, TR = 1.5s, FA = 70, TE = 27.63 ms, FOV = 240, 80, 240 mm, 29 slices ascending, voxel size $= 2.5^3 mm$, slice gap = 0.3 mm, SENSE = 2, 308 volumes) with a 3T Philips Achieva TX MRI scanner with a 32 channel headcoil. Scene presentations (200 ms ON/OFF for 1s) were spaced apart by 10s intervals (in which subjects performed an orthogonal letter task to sustain attention), allowing for reliable estimation of BOLD responses to individual scenes.

The resulting single trial scans were subjected to voxel-wise event-related GLM analysis. This results for each voxel in a beta coefficient, denoting the magnitude of the voxel which are averaged across trials. For each voxel, the local multivariate BOLD response is established using searchlight technique [10] resulting in $\mathbf{y}_k = v_1, ...v_S$ where S denotes the number of voxels within a spherical sphere of radius = 2.5mm, resulting in S(= 27) voxels.

3. VARIATION PARTITIONING USING DISSIMILARITY MATRICES

We use distance-based variation partitioning [11] to study the contribution of HMAX and BoW in explaining fMRI responses to visual scenes.

The variation partitioning [14] algorithm determines the unique contribution of HMAX distance matrix X and BoW distance matrix W in explaining the brain activity distance matrix Y.

First we determined the explained variance of Y by the combination of X and W, $R^2_{Y|X+W}$ done on the basis of the predicted response $\hat{Y}_{X+W}$ resulting from the regression of X and W together on Y. Similarly the fraction of Y explained by X independently based on $\hat{Y}_X$ is $R^2_{Y|X}$ and fraction that is explained by W independently based on $\hat{Y}_W$ is $R^2_{Y|W}$.

The amount of explained variance is then

$$R^2_{Y|X+W} = \frac{trace(\hat{Y}'_{X+W} * \hat{Y}_{X+W})}{trace(Y' * Y)} \quad (1)$$

Similarly the fraction of Y explained by X independently is determined based on $\hat{Y}_X$,

$$R^2_{Y|X} = \frac{trace(\hat{Y}'_{X} * \hat{Y}_{X})}{trace(Y' * Y)} \quad (2)$$

and the fraction that is explained by W independently is given by computing $\hat{Y}_W$,

$$R^2_{Y|W} = \frac{trace(\hat{Y}'_{W} * \hat{Y}_{W})}{trace(Y' * Y)} \quad (3)$$

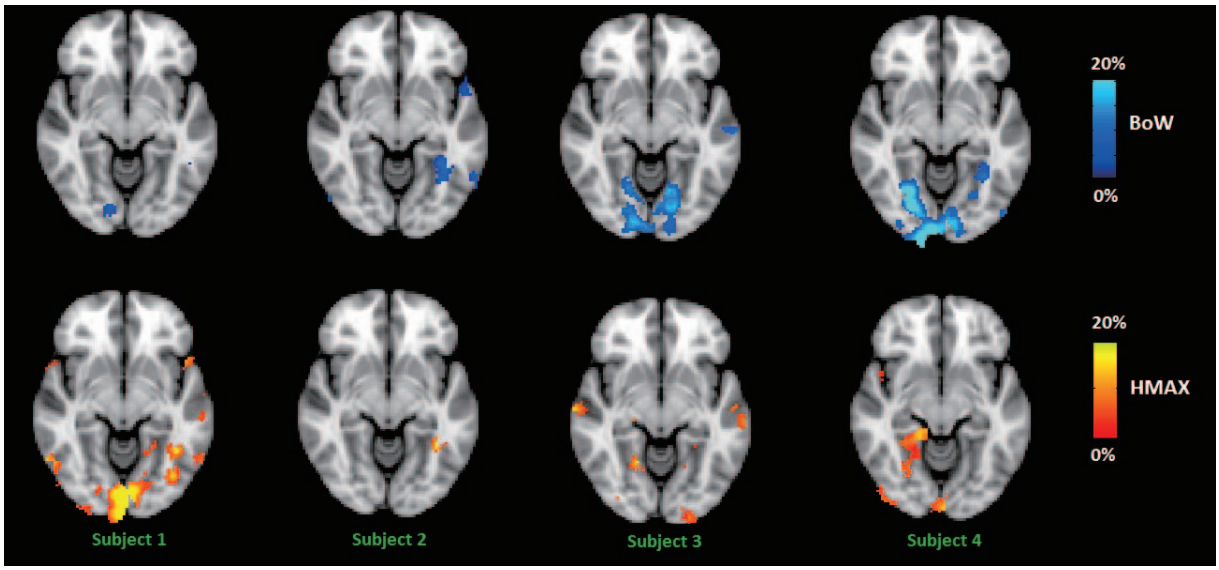

Fig. 1. A. Brain map showing voxelwise the fraction of brain activity explained uniquely by BoW (blue) and HMAX (red) for four subjects.

Unique contributions of HMAX and BoW in explaining local brain activity are computed by subtracting the $R^2_{Y|X}$ from $R^2_{Y|X+W}$ and $R^2_{Y|W}$ from $R^2_{Y|X+W}$ respectively. Note that these statistics are the canonical equivalent of the regression coefficient of determination, R^2 [14].

4. RESULTS

4.1. Subject specific results

We used distance based variation partitioning to determine the unique contribution of HMAX and BoW in explaining fMRI responses. We report only clusters with a minimum of 25 contiguous voxels with significant correlations ($p < 0.05$, ALPHASIM [15]). Figure 1 shows for each subject the amount of explained variance by BoW (blue) and HMAX (red).

BoW also account for brain activity at multiple brain regions, but the regions tend to concentrate in the primary and extrastiate visual cortex areas. In addition, BoW explain up to 18% for the four subjects. Interestingly, for three out of four subjects the maximum explained variances is found in adjacent regions in the primary and extrastriate visual cortex. These results suggest a locus in the visual cortex and consistency across subject for BoW.

HMAX features correlate with brain responses in multiple brain areas. These areas include primary visual cortex (extraction of low-level features), mid-level level areas such as the Lateral occipital cortex (involved in object processing), and higher-level areas such as Parahippocampal gyrus (encoding and recognition of scenes). For the four subjects, the explained variance peaks between 8%-18% and is located at different brain areas (Occipital Pole, Parahippocampal Gyrus, Lateral Occipital Cortex and Lingual gyrus).

4.2. Across subject averages

As averaging fMRI responses across subject may enhance response signals, we repeated our analysis on subject-averaged fMRI responses. As before, figure 2 shows the uniquely explained variances by BoW and HMAX. The maximum explained variance by HMAX is (22%) and by BoW (21%). This peak in explained variance for both the models is interesting, and suggest that visual dictionaries indeed capture information processed in the visual brain. Moreover the peak explained variance by HMAX features occurs slight earlier in the visual hierarchy compared to the area where the peak explained variance by BoW emerges.

Table 1 shows explained variances averaged across predefined regions of interest. The data confirm that the highest explained variances are found in the higher areas TO, V4 and adjacent areas. In addition, the data show that brain activity in the vast majority of voxels in most of the visual regions is accounted for by HMAX, whereas significant voxels in area V1, V2 and V3, are due to BoW (which only account for a small fraction of the the MRI responses). Brain activity in higher level areas such as the Inferiro temporal lobule, Anterior Temporal, and Posterior Temporal is also explained by both the models, but in fewer voxels and to a lesser extent.

Figure 3 shows the explained variance by the combination of the two models (BoW and HMAX) and also the difference in the models (BoW and HMAX) from the combined model. We observe that the peak explained variance is higher for the combined model (dark green) than either of the individual models. The difference between the combined models

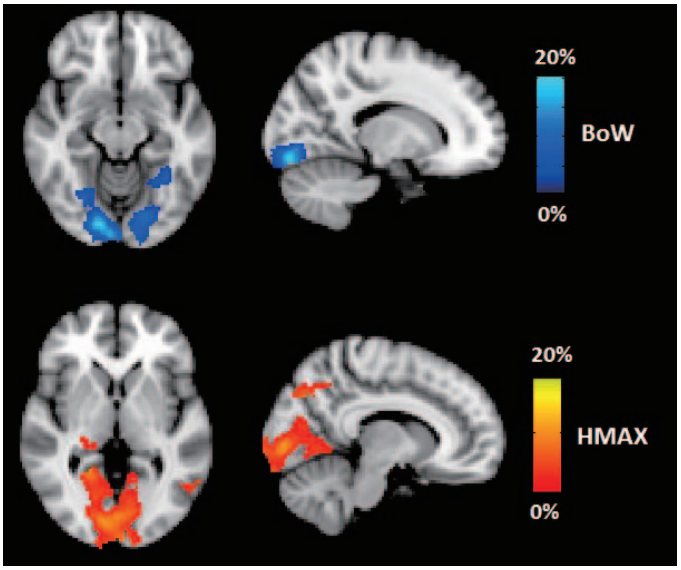

Fig. 2. Visualization of brain activity explained uniquely by BoW (blue) and HMAX features (red) for average subject responses.

Table 1. Number of significant voxels and explained variance for ROI across subject averages

	No of Significant Voxels		Max Explained Variance	
	BoW	HMAX	BoW	HMAX
V4	460	568	21.31	11.42
V123	1359	4002	21.31	22.11
TO	361	741	10.08	21.10
IPL	62	174	5.07	11.15
PT	37	137	4.31	9.54
LO	0	440	0	11.41
AnteriorTemporal	159	180	4.71	7.70
LGN	58	109	4.82	6.23
SPlobule	0	0	0	0
Area5	0	0	0	0
Area7	0	0	0	0

(light green) and the individual models is seen in V1, V2, V3 and V4, suggesting that there is additional value by combining BoW and HMAX.

5. DISCUSSION AND CONCLUSION

The success of models such as HMAX and BoW may be attributed to their use of features of intermediate complexity. Both these models being so widely successful, makes them candidate computational model of intermediate visual processing in the brain. Our results show that in certain brain areas such as V123 and V4 there are overlapping regions in which brain activity is explained by both the models while there are also brain areas where these models explain non-overlapping brain regions. These results provide evidence that both these models adequately capture the visual information that is processed in the brain and the use of visual dictionaries provides a suitable computation for representing intermediate features.

We observe that in the early visual areas and beyond such as the V4, the brain activity is explained in overlapping regions by both these models. This might be since there are similarities in both these models. They both use gradient information, gabor filter in case of HMAX and SIFT in case of BoW, as low level features which explain the sensitivity in the lower visual regions. The overlapping regions in higher brain areas by both these models can be attributed to their use of intermediate features.

The intermediate computation in both these models use the concept of visual dictionaries, albeit constructed differently. But the visual dictonaries in both the models are akin to using medium size image patches that are informative and

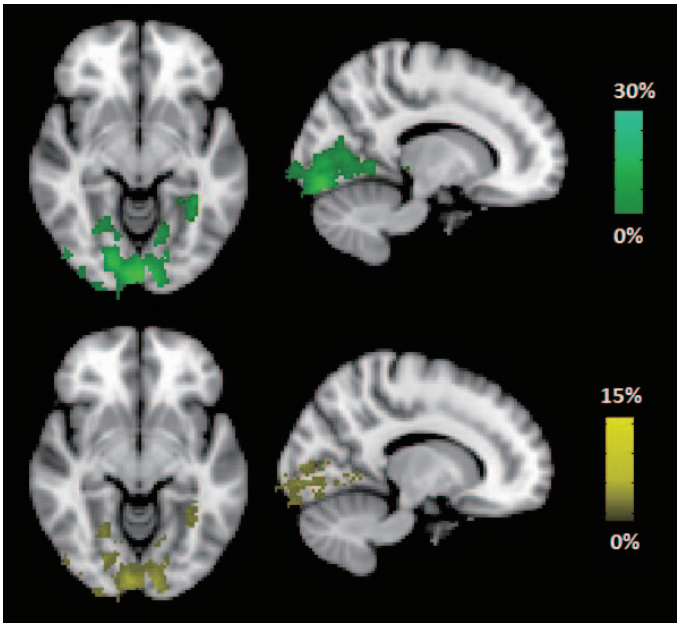

Fig. 3. Visualization of brain activity explained by the combination of the two models (dark green) and the difference in the models(BoW and HMAX) from the combined model (light green) for average subject responses(obtained by subtracting the explained variance of either BoW or HMAX, whichever is higher, from the explained variance of the combined model).

at the same time distinctive. The visual dictionary can be thought of as higher-level visual building blocks composed of slightly larger receptive fields. Thus they are analogous to features of intermediate complexity that play an important role in scene classification. Being compact and rich visual descriptors, the image patches of the visual dictionary may allow for sparse and intermediate representations of objects and scenes. We observe from our results that both these models explain the most brain activity in overlaping higher level visual regions which might be due to the use of visual dictionaries in these models. These results show there is a neural evidence for the similarity in the models and these findings suggest that visual dictionaries might be a computational step to capture the intermediate features of the brain.

There are also differences in how these models explain brain activity. From our results we see that the HMAX model explains more number of voxels and has a higher peak explained variance in most of the brain regions compared to BoW. These two models explain brain activity in certain non overlaping regions. For example while the HMAX model significantly explains brain activity in the LO region, there are no voxels in the LO region explained by BoW. These differences can be attributed to different computations in the two models, for example one is the way visual dictionaries are constructed by HMAX and BoW. In the BoW model, quantization of SIFT patches to the visual words is done for scene representation. This represents what one can say quantization error, i.e, certain amount of image information is lost while assigning the patch to a cluster. However in the HMAX the elements of the visual dictionary are convolved over the image from the second layer and is fed to the next layer. Thus one or more differences in the computation of visual dictionary might reflect in the differences in explained brain activity of HMAX and BoW.

In conclusion while there are certain brain regions explained by both HMAX and BoW, there are regions associated with visual processing not explained by the models. These regions mostly are in the higher brain areas. The next step would be to understand what are the computations required for the models to explain brain activity in the higher brain regions significantly. This might helps us answer question on how best to compute intermediate representations and also to use higher level features which is still a open question. Thus going forward, these higher computation steps could derive inspiration from the brain. In the future we will also address other open questions such as the role of dimensionality of visual dictionary and the role of quantization error. More generally, it is interesting to study the repeatability of our results when creating visual dictionary based on other data sets and when using fMRI responses to a wider range of natural scenes.

Acknowledgement

This research was supported by the Dutch national public-private research program COMMIT.

6. REFERENCES

[1] Field DJ, Relations between the statistics of natural images and the response properties of cortical cells. J Opt Soc Am A 4: 23792394 1987.

[2] Olshausen BA and Field DJ, Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature 381, 607610 1996.

[3] Karklin Y and Lewicki MS, Emergence of complex cell properties by learning to generalize in natural scenes. Nature 457, 8386 2009.

[4] Ullman S, Vidal-Naquet M and Sali E, Visual features of intermediate complexity and their use in classification. Nature Neuroscience. 5, 682687 2002.

[5] Riesenhuber. M and Poggio. T, Hierarchical Models of Object Recognition in Cortex. Nature Neuroscience 2: 1019-1025 1999.

[6] J. Philbin, O. Chum, M. Isard, J. Sivic and A. Zisserman, Object retrieval with large vocabularies and fast spatial matching. IEEE Computer Society Conference on Computer Vision and Pattern Recognition(CVPR) 2007.

[7] A. F. Smeaton, P. Over, and W. Kraaij, Evaluation campaigns and trecvid. MIR 2006.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, The pascal visual object classes (VOC) challenge. IJCV 2010.

[9] D. Parikh and C.L. Zitnick, The Role of Features, Algorithms and Data in Visual Recognition. Proc. IEEE Conf. Computer Vision and Pattern Recognition(CVPR) 2010.

[10] N. Kriegeskorte, M. Mur and P. Bandettini, Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers System Neuroscience 2, p.4 2008.

[11] Legendre, P. and M. J. Anderson, Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. Ecological Monographs 69: 1 - 24 1999.

[12] Lowe, D. G. Distinctive Image Features from Scale Invariant Keypoints. International Journal of Computer Vision, 60(2),91 - 110 2004.

[13] Wichmann FA Person, Braun DI and Gegenfurtner, KR Person Phase noise and the classification of natural images. Vision Research 46(8-9), 1520-1529 2006.

[14] Peres-Neto P. R, Legendre P, Dray S and Borcard D, Variation partitioning of species data matrices: estimation and comparison of fractions. Ecology 87: 2614 - 2625 2006.

[15] Cox, R.W. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. International Journal of Computers and biomedical research, 29(3), 16273 1996.