The Elastic Net as Visual Category Representation: Visualisation and Classification

Dror Cohen and Andrew P. Papliński

Clayton School of IT, Monash University, Australia
{Andrew.Paplinski,Dror.Cohen}@monash.edu

Abstract. In this paper we use the Elastic Net (EN) [9] as a visual category representation in feature space. We do this by training the EN on the high dimensional Pyramid Histogram of Visual Words (PHOW) features [2] often used in modern visual categorisation. By employing the topography preserving properties of the EN we visualise the features and draw some novel conclusions. We demonstrate how the EN can also be used as a Region of Interest detector [1]. Finally, inspired by biological vision we propose a new Visual Categorisation scheme that uses ENs as visual category representations. Our method shows promising results when tested on the Caltech101 [12] data set with several interesting future directions.

Keywords: Elastic Net, Visual Categorisation, Object Recognition, Caltech101.

1 Introduction

In this paper we use the EN [9], a type of probabilistic self-organising map, to form a visual category representation in the high dimensional Pyramid Histogram of Visual Words (PHOW) [2] feature space frequently used in modern Visual Categorization systems [5]. Employing the topography preserving properties of the EN we are able to visualise the PHOW features and draw a number of novel conclusions. We also demonstrate the potential of using the EN as a [2] detector. We then continue to suggest a novel recognition framework that uses ENs as category representations in the PHOW feature space.

The task of Visual Categorisation (assigning an image to a category by processing the image) remains one of the principle problems of Computer Vision. In recent years, a general recognition scheme has emerged that is able to produce good results in challenging data sets such as Caltech101 [12] or the PASCAL Visual Object Class Challenge [11]. Central to this scheme is the representation of image patches by robust and invariant features. The high dimensionality of these features makes interpreting them in an intuitive way difficult. As the following sections will demonstrate, using the EN as a category representation in feature space we are able to asses these features, comment on their parameters and also compare images in a intuitive and useful way.

T. Huang et al. (Eds.): ICONIP 2012, Part II, LNCS 7664, pp. 133–140, 2012.

The modern approach to visual categorisation can be broadly divided into the following four steps: extraction of local features (e.g. [20,2]), construction of characteristic codebook, encoding of an image descriptor (see [5,7]), and training of a classifier (e.g. SVM, Naive Bayes).

As the system's performance depends on all of the above steps, comparison of different schemes is difficult. This issue has been rectified to some extent in [5] where different encoding methods have been compared, keeping the type of features and codebook size consistent. Inline with this comparison, we have chosen to study the PHOW features. Further we will use the VLFeat library [23] to compute these features, maintaining the default settings as in [5].

1.1 Elastic Net

The Elastic Net (EN) was first developed as an analogue approach to the Traveling Salesman Problem [9], and is well known as a model for activity dependent plasticity in V1 [13].

In the EN, nodes are evenly spaced on a latent space (usually 2D) lattice. Each node m has a position vector $v_m \in \mathbb{R}^2$ and a weight vector $w_m \in \mathbb{R}^D$. A set of observation $X = \{x_1, x_2, ..., x_N\}, x_n \in \mathbb{R}^D$, is introduced for which the weights are updated according to the following rule:

$$\Delta w_m = \sum_n \rho_{nm}(w_m - x_n) + \alpha k \sum_{j \in \Lambda_m} |w_m - w_j|^2 \tag{1}$$

Here Λ_m represents the set of weights in the 'lattice neighbourhood' of node m, and α and k are parameters. Typically, Λ_m consists of the nodes directly South, North, East and West from node m and this is the arrangement we employ here. The terms ρ_{nm} are the 'responsibilites' given by

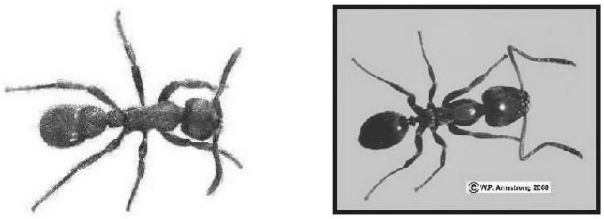
$$\rho_{nm} = \frac{\Theta(x_n|w_m, k)}{\sum_s \Theta(x_n|w_s, k)}, \quad \Theta(x_n|w_m, k) = \exp(\frac{-|x_n - w_m|^2}{2k^2}) \tag{2}$$

An EM type algorithm [8] is employed where at each iteration the responsibilities (2) are calculated, followed by the update equation (1) and consequent recalculation of the responsibilities. Throughout this process the parameter k is gradually reduced so that the weight of the 'continuity' term (second term in (1)) compared to the 'coverage' term (first term in (1)) is reduced.

In a probabilistic formulation the EN can be interpreted as a *maximum a posteriori* estimate for a Gaussian Mixture Model (GMM) with a prior over the weights [10]. Detailed analysis of the EN and the properties of this prior can be found in [4]. Our choice of the EN over other topography preserving maps is based on our previous work with this model [6], its biological relevance and its probabilistic interpretation. Throughout this work we only consider the case where the latent space is two dimensional and may thus refer to the EN as a 'map'.

2 Using the EN as a Visual Category Representation

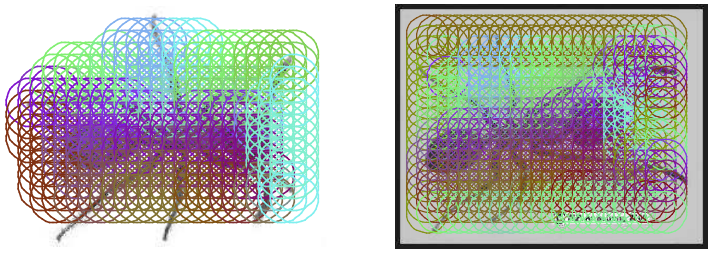
We form a visual category representation in the high dimensional PHOW features by training an EN on the concatenated features from a set of training images of a given category. To visualise this representation we select one training image and one test image and apply their features to the EN. For each feature we plot the mean location on the 2D map, given by $u_n = \sum_m \rho_{nm} v_m$. To improve the visualisation we assign a colour to each feature by transforming the mean location to the green and blue components, keeping the red component fixed. We use the same colour to mark the feature's patch on to the original image. We demonstrate this on the ant category from the Caltech101 data set. To train the map, 15 images are chosen in random. Fig. 1a and b show the training and test images we used to visualise.



(a) (b)

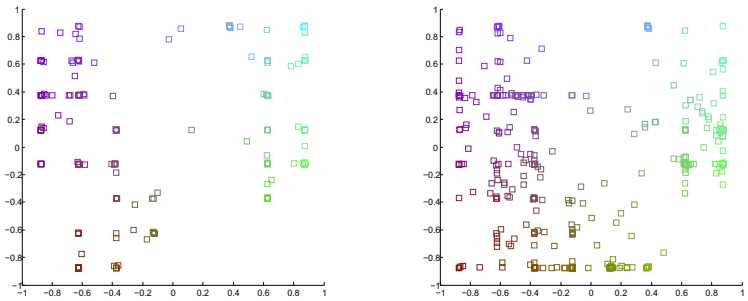
Fig. 1. Images from the ant class in Caltech101. (a) An image used in training the EN. (b) An image used for testing.

Fig. 2a and b show the training and test images with the patches superimposed and coloured as described. A MATLAB [17] version of the EN is available from [3]. Fig. 3a and b show the mean locations for the training and test image respectively. The gradual shifts in colour seen in Fig. 2 corresponds to nearby features (features from adjacent patches) being mapped to nearby nodes on the map. The visualisation suggests that strongly overlapping patches will result in



(a) (b)

Fig. 2. Visualising the PHOW features. The feature patch is marked by a circle and coloured according to the mean location.



(a) (b)

Fig. 3. Mean locations of the training (a) and testing (b) images' features on the trained EN

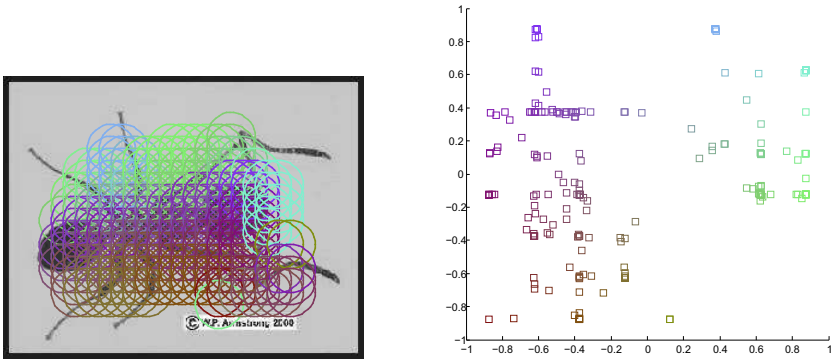
a feature data set that contains duplicate (or very similar) features. This may unduly increase the computational time in the codebook construction and image encoding stages (section 1). In the case where features are extracted over a few different scales (spatial bin of the PHOW feature), the larger scale features will contain more duplicates (as the patches will overlap more). This may skew the quantisation stage towards the possibly over represented, larger scale features.

If many duplicates are present than an an incremental rather than batch type quantisation approach may converge faster. To reduce redundancy in the data set the larger scale features can be evaluated at larger step sizes, so that there is no greater patch overlap in the larger scale features.

The high dimensional features are not directly interoperable as more intuitive concepts such as "ant rear leg". Using the visualisations we can identify such concepts. For example, for the images in Fig. 1 (which were intentionally chosen as similar) we may expect to find similar colours and colour shifts in similar parts of the ant. This is observed to some degree in Fig. 2. For example, the legs in both images have the same colour and colour shifts. This is less clear in the area corresponding to the ant's antennas, which differs between the ant images (see Fig. 1).

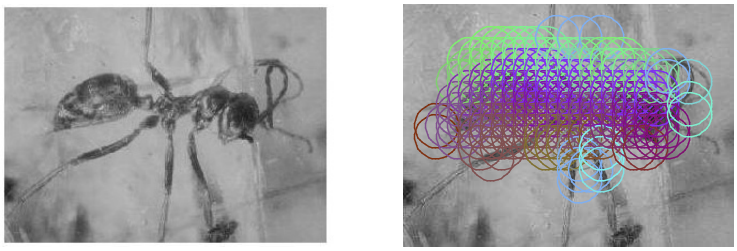
We can improve the correspondence between the images by rejecting features that are not well represent in the map. We do this by thresholding according to the 'posterior entropy' $pH_n = \sum_m log(\rho_{nm})\rho_{nm}$, which provides a measure of how peaked the posterior distribution is across the map. In Fig. 4a we removed all the features with posterior entropy below the median for the image. Similarly, Fig. 4b shows the remaining mean locations. The correspondence with the training image is now clearer.

It can be seen in Fig. 4a that features corresponding to the 'copywrite' symbol are removed by the posterior entropy thresholding. By filtering in this way we may use the map as a Region of Interest Detector. We demonstrate this on an ant image with a more challenging background in Fig. 5.



(a) (b)

Fig. 4. Visualising the PHOW features by thresholding the posterior entropy (a) Visualisation after thresholding the posterior entropy. (b) Mean locations of the features after thresholding.



(a) (b)

Fig. 5. Using the EN as a Region of Interest Detector

3 New Recognition Framework

In this section we demonstrate that the category representations formed by ENs can be used in visual categorisation. Our approach draws on biological vision where there is some evidence that distinct region of the cortex represent high level categories such as faces, animals etc. [14]. Given that the EN is an established model for plasticity in V1 [13], we motivate the idea that a similar process may also occur in higher visual processing areas.

First we extract the PHOW features from a number of training images for each category. The features are concatenated together to form the training data set for an EN. That is, each category is represented by an EN. In the testing stage, the test image's features are provided to each map. The image is then assigned to the map with the highest log-likelihood. This scheme is demonstrated in Fig. 6. Our approach is considerably simpler than that described in section 1, doing away with both the encoding and the classification stages. Further, since a classifier is not employed, categories can be added on the fly by adding another map.

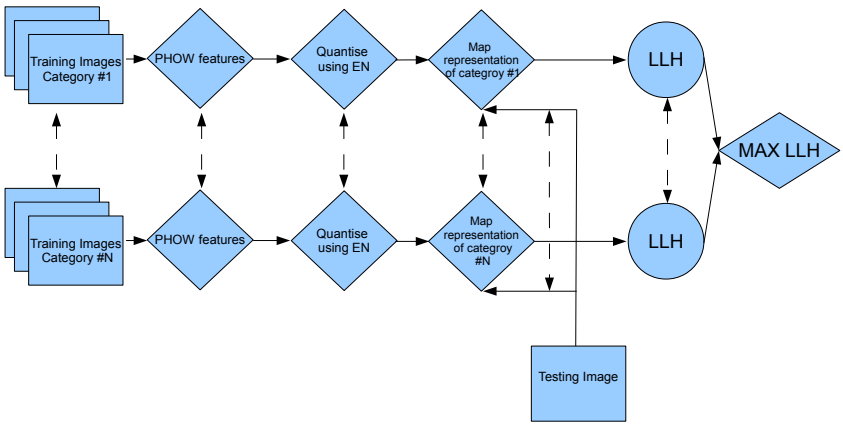


Fig. 6. New recognition framework

To the best of our knowledge this is the first application of ENs combined with PHOW features for visual categorisation. Related work in facial recognition and content based image retrieval can be found in [16] and [21].

3.1 Experimental Setup

We evaluate the system on the Caltech101 data set. The only pre-processing we preform is to scale the features so that the the collection of features describing each image has components in the range $[0-1]$. The model parameters for the EN are m, α the initial and final values of k (k_{in} and k_{end}) and the number of iterations (which determines the rate of annealing). Using the visualisation from the previous section and with some further experimentation we found that the following values yield good results. $m = 64, \alpha = 10, k_{in} = 1, k_{end} = 0.1, iter = 1000$. We use this setup for all maps.

Table 1 shows our results for different step and bin sizes of the PHOW features. The results are obtained by averaging three runs with 15 random training images and up to random 30 testing images (some categories have less than 45 images) for each category.

As can be seen from Table 1 the performance varies significantly with the choice of spatial bin size. To account for this scale variability we add the log-likelihood from each corresponding map and use this new, cumulative quantity

Table 1. Average recognition rate for different PHOW Step and Spatial Bin sizes

Step Size, Spatial Bin Size	Average recognition rate
[4,4]	34.56 ± 1.1
[5,10]	45.15 ± 1.2
[10,20]	50.34 ± 0.5
[15,30]	47.48 ± 0.8
Cumulative	54.60 ± 1.0

as our decision criteria. The results for this approach are shown on the last row of Table 1 and improves the performance by a further ∼4.5%

These results are comparable with earlier attempts at visual categorisation of this data set [20,18]. However, still ∼10% shy from the state of the art (single feature) models which achieve ∼64% [23].

3.2 Discussion

There are several modifications that will improve the system's performance. Firstly, the EN can be extended to include arbitrary covariance rather then the uni-variance used. Also, the parameters for each map can be optimised separately through Bayesian estimation of hyper parameters [22]. A Variational treatment similar to that in [19] is also possible. Our approach did not use any spatial binning to capture geometric information [15,1] which will likely improve our performance.

Since we train a map for each category it is possible for maps to share features. For example, if we assume that we are able to correctly discriminate between the top three maps (based on the cumulative LLH) than our performance improves to ∼68%. One way of addressing this would be to employ a hierarchical structure. Another could be to train the EN in a more discriminative manner by, for example, considering the other ENs nodes when computing the responsibilities. It seems reasonable to assume that with some of these suggested improvements our method will rival the state of the art.

Our approach also offers some advantages. Firstly, it is both simple and transparent in the sense that each map represents a category. Secondly, categories can be added on the fly by simply training another map. Different types of features can easily be incorporated by training another EN. Our approach is also parallelisable since each map can be trained and tested independently of the other maps.

4 Conclusion

In this paper we used the Elastic Net as a visual category representation in feature space. Using this representation we visualised the high dimensional features and drew some novel conclusions about the PHOW feature parameters. The visualisation also allowed interpretation of the features in terms of higher level concepts. We demonstrated that the EN could be used as a Region of Interest detector by rejecting features that were not well represented by the map. We proposed a new recognition scheme that uses ENs as visual categories and demonstrated its validity. We suggested a number of possible improvements that would make our proposed approach competitive with the state of the art, while maintaining some key advantages such as simplicity and being able to add categories on the fly.

References

1. Bosch, A., Zisserman, A., Munoz, X.: Representing shape with a spatial pyramid kernel. Image Processing 5(2), 401–408 (2007)
2. Bosch, A., Zisserman, A., Munoz, X.: Image Classification using Random Forests and Ferns. In: Proc. ICCV, vol. 21, pp. 1–8 (2007)

3. Carreira-Perpinan, M.: Generalised elastic nets (2003), http://faculty.ucmerced.edu/mcarreira-perpinan/papers.html
4. Carreira-Perpiñán, M.Á., Dayan, P., Goodhill, G.J.: Differential Priors for Elastic Nets. In: Gallagher, M., Hogan, J.P., Maire, F. (eds.) IDEAL 2005. LNCS, vol. 3578, pp. 335–342. Springer, Heidelberg (2005)
5. Chatfield, K., Lempitsky, V., Vedaldi, A., Zisserman, A.: The devil is in the details: an evaluation of recent feature encoding methods. In: Proc. BMVC 2011, pp. 1–12 (2011)
6. Cohen, D., Papliński, A.P.: A comparative evaluation of the Generative Topographic Mapping and the Elastic Net for the formation of Ocular Dominance stripes. In: Proc. WCCI–IJCNN, pp. 3237–3244. IEEE (2012)
7. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. Earth 1, 22 (2004)
8. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum Likelihood for Incomplete Data via the EM Algorithm. J. Royal Statistical Society B 39(1), 1–38 (1977)
9. Durbin, R., Willshaw, D.: An analogue approach to the travelling salesman problem using an elastic net method. Nature 326(6114), 689–691 (1987)
10. Durbin, R., Szeliski, R., Yuille, A.: An analysis of the Elastic Net Approach to the Traveling Salesman Problem. Neural Computation 1(3), 348–358 (1989)
11. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL Visual Object Classes (VOC) Challenge. Int. J. Computer Vision 88(2), 303–338 (2010)
12. Fei-Fei, L., Fergus, R., Perona, P.: Learning Generative Visual Models from Few Training Examples. In: Proc. CVPR (2004, 2008)
13. Goodhill, G.J.: Contributions of theoretical modeling to the understanding of neural map development. Neuron 56(2), 301–311 (2007)
14. Gross, C.G.: Coding for visual categories in the human brain. Nature Neuroscience 3(9), 855–856 (2000)
15. Lazebnik, S., Schmid, C., Ponce, J.: Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In: Proc. CVPR, vol. 2, pp. 2169–2178 (2006)
16. Lefebvre, G., Garcia, C.: A probabilistic Self-Organizing Map for facial recognition. In: Proc. ICPR, pp. 1–4 (2008)
17. MATLAB: R2012a, The MathWorks Inc. (2012)
18. Mutch, J., Lowe, D.G.: Multiclass Object Recognition with Sparse, Localized Features. In: Proc. CVPR, vol. 1, pp. 11–18 (2006)
19. Olier, I., Vellido, A.: Variational Bayesian Generative Topographic Mapping. J. Mathematical Modelling and Algorithms 7, 371–387 (2008)
20. Serre, T., Wolf, L., Poggio, T.: Object Recognition with Features Inspired by Visual Cortex. In: Proc. CVPR, vol. 2, pp. 994–1000 (2005)
21. Sfikas, G., Constantinopoulos, C., Likas, A., Galatsanos, N.P.: An analytic distance metric for Gaussian mixture models with application in image retrieval. Framework, 835–840 (2005)
22. Utsugi, A.: Density estimation by mixture models with smoothing priors. Neural Computation 10(8), 2115–2135 (1998)
23. Vedaldi, A., Fulkerson, B.: VLFeat: An Open and Portable Library of Computer Vision Algorithms (2008), http://www.vlfeat.org/