Weakly Supervised Learning of Foreground-Background Segmentation using Masked RBMs

Nicolas Heess, Nicolas Le Roux, John Winn

Weakly Supervised Learning of Foreground-Background Segmentation using Masked RBMs

Nicolas Heess[1], Nicolas Le Roux[2], and John Winn[3]

[1] University of Edinburgh, IANC, Edinburgh, UK
[2] INRIA, Sierra Team, Paris, France
[3] Microsoft Research, Cambridge, UK

Abstract. We propose an extension of the Restricted Boltzmann Machine (RBM) that allows the joint shape and appearance of foreground objects in cluttered images to be modeled independently of the background. We present a learning scheme that learns this representation directly from cluttered images with only very weak supervision. The model generates plausible samples and performs foreground-background segmentation. We demonstrate that representing foreground objects independently of the background can be beneficial in recognition tasks.

Keywords: RBM, segmentation, weakly supervised learning

1 Introduction

Learning generative models of natural images is a long-standing challenge. Recently, a new spectrum of approaches, loosely referred to as "deep learning" (DL), has led to advances in several AI-style learning tasks. At the heart of this framework is the use of RBMs for greedy learning of multi-layered representations such as Deep Belief Networks (DBN, [2]) and Deep Boltzmann Machines (DBM, [7]). However, despite interesting applications in vision (e.g. [3]), it has become apparent that the basic formulation of the RBM is too limited to model images well, and several alternative formulations have recently been proposed (e.g. [6]). One powerful notion from the computer vision literature is that of a layered representation. This allows images to be composed from several independent objects and can account for occlusion (e.g. [10, 11]). In [4], we have proposed a model that introduces such a layered representation into the DL framework. In this model, the Masked RBM (MRBM), an image is composed of several regions, each of which is modeled in terms of its shape and appearance. The region shape determines where a region is visible, and the appearance determines the color or texture of the region, while overlapping regions occlude each other. We used this architecture to formulate a generative model of lower-level structure in *generic* images and therefore assumed that all regions were equivalent, i.e. all regions were governed by the same shape and appearance models, and that shape and appearance were independent. For higher-level tasks such as recognition, however, the different regions of an image are not equivalent and some

are more interesting than others. Here, we show that separating an image into *qualitatively different* layers and *jointly* modeling shape and appearance allows us to learn and represent specific objects or object categories *independently* of the background. As a result, the representation of the foreground is less affected by background clutter. Our model is able to perform foreground-background segmentation, and it can generate new instances of the foreground objects (shape and appearance). In particular, we show that learning is possible directly from cluttered images and requires only very weak supervision: inspired by [11], we bootstrap learning with an approximate model of the background. This is easily obtained by training on general natural images and sufficient for learning then to proceed without further supervision: foreground objects can be detected as outliers under the background model and a model of the foreground can thus be learned from the regularities of these outliers across training images. To our knowledge, foreground-background segmentation has not previously been addressed in the DL framework. Tang [8] proposes a model that is related to ours and applies it to the problem of recognition under occlusion, but considers a simpler scenario with binary images and fully supervised learning only.

2   Model

Our model extends the MRBM presented in [4]: instead of modeling general images that consist of generic and equivalent regions, it assumes that images contain a single foreground object in front of a cluttered background (which can and often will also contain parts of other objects, but these are not explicitly modeled). Foreground and background are assumed to be independent and the background is occluded by the foreground object. In the model, this is achieved by composing the *observed* image from two *latent* images: one for the background, and one for the foreground. The background image is visible only where the foreground is not, and the visibility of the foreground image is determined by a binary mask. Intuitively speaking, the foreground image determines the *appearance* of the foreground object, and the mask determines its *shape*. The model is a pixel-wise binary mixture with the mixture component for each pixel specified by the mask. Denoting the observed image by $\mathbf{x}$, the background image by $\boldsymbol{v}^{\mathrm{B}}$, and the appearance and shape of the foreground by $\boldsymbol{v}^{\mathrm{F}}$ and $\mathbf{m}$ respectively, the model can be written as

$$P(\mathbf{x}){=}\sum_{\mathbf{m}}\int\!\mathrm{d}\boldsymbol{v}^{\mathrm{B}}\!\int\!\mathrm{d}\boldsymbol{v}^{\mathrm{F}}\!\left(\prod_i\delta\big[v_i^{\mathrm{F}}{=}x_i\big]^{m_i}\delta\big[v_i^{\mathrm{B}}{=}x_i\big]^{(1-m_i)}\right)\!p_{\mathrm{FG}}(\boldsymbol{v}^{\mathrm{F}},\mathbf{m})p_{\mathrm{BG}}(\boldsymbol{v}^{\mathrm{B}}) \quad (1)$$

where i is the pixel index, $m_i \in \{0, 1\}$, and the product of delta functions forces, for each pixel, one of the two latent images to take on the value of the observed image at that pixel. The mask determines which latent image is chosen. We formulate the priors over the background image and over foreground appearance and shape as RBMs. Assuming that pixels are continuous valued in $[0, 1]$, we choose for p_{BG} a special form of the Beta RBM with energy $E_{\mathrm{Beta}}(\boldsymbol{v}^{\mathrm{B}}, \mathbf{h})$ described in [4] so that $p_{\mathrm{BG}}(\boldsymbol{v}^{\mathrm{B}}) = 1/Z \sum_{\mathbf{h}} \exp\{-E_{\mathrm{Beta}}(\boldsymbol{v}^{\mathrm{B}}, \mathbf{h}; \Theta_{\mathrm{BG}})\}$. Unlike the

Gaussian RBM with fixed variance commonly used for continuous data (e.g. [3]), the Beta RBM models mean *and* variance of the visible units.

The model of the foreground object defines a joint distribution over a continuous valued image and a binary mask. Here, we are interested in the case where shape and appearance are *dependent* and we will therefore model them jointly. This is in contrast with the approach taken in [4] where they were assumed to be independent, i.e. $p_{\mathrm{FG}}(\boldsymbol{v}^{\mathrm{F}}, \mathbf{m}) = p_{\mathrm{FG}}(\boldsymbol{v}^{\mathrm{F}})p_{\mathrm{FG}}(\mathbf{m})$. Thus, in this new model, we use a particular form of the RBM which has two sets of visible units, a set of binary units for the mask $\mathbf{m}$ and a set of continuous valued units for the appearance image $\boldsymbol{v}^{\mathrm{F}}$:

$$E_{\mathrm{mixed}}(\mathbf{v}, \mathbf{m}, \mathbf{h}; \Theta) = E_{\mathrm{Bin}}(\mathbf{m}, \mathbf{h}; \Theta^{\mathrm{S}}) + E_{\mathrm{Beta}}(\mathbf{v}, \mathbf{h}; \Theta^{\mathrm{A}}) \quad , \tag{2}$$

where $E_{\mathrm{Bin}}(\mathbf{m}, \mathbf{h}; \Theta) = \mathbf{m}^T W \mathbf{h} + \mathbf{b}^T \mathbf{m}$ is the energy function of a binary RBM, and the joint distribution is thus given by: $p_{\mathrm{FG}}(\boldsymbol{v}^{\mathrm{F}}, \mathbf{m}) = \frac{1}{Z} \sum_{\mathbf{h}} \exp\{-E_{\mathrm{mixed}}(\boldsymbol{v}^{\mathrm{F}}, \mathbf{m}, \mathbf{h}; \Theta)\}$.

Inference: although exact inference is intractable, an efficient Gibbs sampling scheme exists, as detailed in [4]. Let $\boldsymbol{h}^{\mathrm{F}}$ and $\boldsymbol{h}^{\mathrm{B}}$ denote the hidden units of the foreground and the background model respectively, then the following properties admit a Gibbs sampling scheme in which the three sets of variables $(\boldsymbol{h}^{\mathrm{F}}, \boldsymbol{h}^{\mathrm{B}})$, $\mathbf{m}$, and $(\boldsymbol{v}^{\mathrm{F}}, \boldsymbol{v}^{\mathrm{B}})$ are sampled in turn:

1. given $\boldsymbol{v}^{\mathrm{F}}$ and $\boldsymbol{v}^{\mathrm{B}}$ the hidden units $\boldsymbol{h}^{\mathrm{F}}$ and $\boldsymbol{h}^{\mathrm{B}}$ are conditionally independent, i.e. $p(\boldsymbol{h}^{\mathrm{F}}|\boldsymbol{v}^{\mathrm{F}}) = \prod_j p(h_j^{\mathrm{F}}|\boldsymbol{v}^{\mathrm{F}})$ and $p(\boldsymbol{h}^{\mathrm{B}}|\boldsymbol{v}^{\mathrm{B}}) = \prod_j p(h_j^{\mathrm{B}}|\boldsymbol{v}^{\mathrm{B}})$,
2. given the hidden variables $\boldsymbol{h}^{\mathrm{F}}$, $\boldsymbol{h}^{\mathrm{B}}$ and the image $\mathbf{x}$, the variables of the mask, foreground image, and background image are pixel-wise conditionally independent, i.e. $p(\boldsymbol{v}^{\mathrm{F}}, \boldsymbol{v}^{\mathrm{B}}, \mathbf{m}|\boldsymbol{h}^{\mathrm{F}}, \boldsymbol{h}^{\mathrm{B}}, \mathbf{x}) = \prod_i p(v_i^{\mathrm{F}}, v_i^{\mathrm{B}}, m_i|\boldsymbol{h}^{\mathrm{F}}, \boldsymbol{h}^{\mathrm{B}}, \mathbf{x})$,
3. $p(v_i^{\mathrm{F}}, v_i^{\mathrm{B}}, m_i|\boldsymbol{h}^{\mathrm{F}}, \boldsymbol{h}^{\mathrm{B}}, \mathbf{x})$ can be decomposed as follows

$$p(v_i^{\mathrm{F}}, v_i^{\mathrm{B}}, m_i|\boldsymbol{h}^{\mathrm{F}}, \boldsymbol{h}^{\mathrm{B}}, \mathbf{x}) = p(v_i^{\mathrm{F}}, v_i^{\mathrm{B}}|m_i, \boldsymbol{h}^{\mathrm{F}}, \boldsymbol{h}^{\mathrm{B}}, \mathbf{x})p(m_i|\boldsymbol{h}^{\mathrm{F}}, \boldsymbol{h}^{\mathrm{B}}, \mathbf{x}) \tag{3}$$

$$p(m_i = 1|\boldsymbol{h}^{\mathrm{F}}, \boldsymbol{h}^{\mathrm{B}}, \mathbf{x}) = \frac{p_{\mathrm{FG}}(v_i^{\mathrm{F}} = x_i, m_i = 1|\boldsymbol{h}^{\mathrm{F}})}{p_{\mathrm{BG}}(v_i^{\mathrm{B}} = x_i|\boldsymbol{h}^{\mathrm{B}})p_{\mathrm{FG}}(m_i = 0|\boldsymbol{h}^{\mathrm{F}}) + p_{\mathrm{FG}}(v_i^{\mathrm{F}} = x_i, m_i = 1|\boldsymbol{h}^{\mathrm{F}})} \tag{4}$$

$$p(v_i^{\mathrm{F}}, v_i^{\mathrm{B}}|m_i, \boldsymbol{h}^{\mathrm{F}}, \boldsymbol{h}^{\mathrm{B}}, \mathbf{x}) = \begin{cases} \delta\left[v_i^{\mathrm{F}} = x_i\right] p_{\mathrm{BG}}(v_i^{\mathrm{B}}|\boldsymbol{h}^{\mathrm{B}}) & \text{if } m_i = 1 \\ \delta\left[v_i^{\mathrm{B}} = x_i\right] p_{\mathrm{FG}}(v_i^{\mathrm{F}}|\boldsymbol{h}^{\mathrm{F}}) & \text{otherwise.} \end{cases} \tag{5}$$

Learning: during learning with unlabeled data only $\mathbf{x}$ is observed. We use an EM-like approach in which inference of $\boldsymbol{v}^{\mathrm{F}}$, $\boldsymbol{v}^{\mathrm{B}}$, and $\mathbf{m}$ alternates with updates of the model parameters. Once these variables have been inferred, they can be treated as "observed" data for one of the usual learning schemes for RBMs such as contrastive divergence (CD, [1]) or stochastic maximum likelihood (SML, also referred to as "persistent CD" [9]), which we use in the experiments below. SML relies on persistent chains of samples to represent the model distribution which are updated by one step of Gibbs sampling per iteration. Note that, due to the directed nature of the mixture in (1), the persistent Markov chains representing the model distributions of the two RBMs for foreground and background do not interact, i.e. we run independent chains as if we were training both RBMs separately. Fully unsupervised learning is possible in principle but likely to be very

difficult for all but very simple datasets. We therefore consider a "weakly super-vised" scenario related to [11]: we assume that we have some general knowledge of the statistical regularities to be expected in the *background*. This approximate model of the background can be obtained by training on general natural images and it allows us to bootstrap learning of the foreground model from un-segmented training data. Foreground objects stand out from the background, i.e. they appear as "outliers" to the background model which forces them to be explained by the foreground model. Although this detection is initially un-reliable, the foreground model can then learn about the consistencies of these outliers across the training images which eventually leads to a good model of the foreground objects without any additional information being provided or prior knowledge (e.g. coherence or convexity) being used.

3   Experiments

Datasets & evaluation: we evaluate the model and the learning scheme on two datasets: a toy dataset and a more challenging, modified version of the "Labeled faces in the wild-A" (LFW-A)-dataset [5, 12]. The toy data consist of 16×16 pixel image patches that contain two classes of foreground "objects" against backgrounds that are randomly cropped patches from the VOC 2009 dataset. The two classes differ in appearance and shape (i.e. shape and appearance are *dependent*) and objects can appear at various positions in the patch (see Fig. 1). For the LFW-A dataset, the original images of size 250×250 were cropped to 210×210 pixels and down-scaled to 32×32 pixels. We used the first 13000 im-ages in the dataset for training and the remaining 233 images for test purposes. Examples are shown in Fig. 2a. We evaluate the quality of learned models for both datasets by sampling[4] and by performing FG-BG segmentation on unseen test images[5]. We also investigate whether the FG-BG model's ability to model the object of interest independently of the background provides additional ro-bustness of the latent representation to background clutter in a recognition task.

    Models & learning: for our weakly supervised learning scheme, we first learned models of the background by training Beta RBMs on large sets of natural image patches using SML. We next trained the foreground RBMs in the context of the full model (eq. 1 as described in section 2). For each training data point, we store the state of the latent variables $\boldsymbol{v}^{\mathrm{F}}$, $\boldsymbol{v}^{\mathrm{B}}$, and $\mathbf{m}$ from one epoch to the next. We update them by one step of Gibbs sampling before using the state of the latent variables for the current batch and the particles in the persistent chains for the foreground model to compute the gradient step for the parameters of the foreground RBM in the usual manner [9]. Fig. 2e illustrates how the mask $\mathbf{m}$ converges for a particular *training* image *during learning* over 1000 epochs. For the model to converge well on the face dataset we further initialized the weights

---

[4] Markov chains are initialized with random noise; we use the conditional means of the visible units given the hidden units in the final step for visualization.

[5] We provide a comparison with the performance of a simple conditional random field (CRF) in the supplemental material [13].
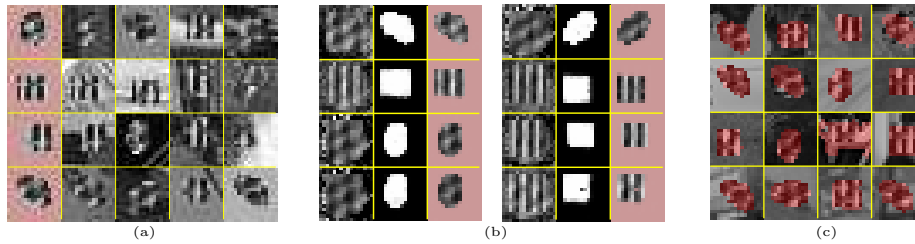
(a)   (b)   (c)

Fig. 1. *(a)* Training toy data: The two classes of "objects" are (1) rectangles of 9 different sizes and (2) four kinds of round shapes. Objects can appear at different locations in the image and are filled with two different types of textures. The texture varies from training image to training image but rectangles are always filled with one type of texture and round objects with the other. The first columns shows "ground truth", i.e. objects in isolation. The remaining columns show actual training data, i.e. objects embedded in natural image backgrounds. *(b)* Samples from the model: In each block the left column shows $\boldsymbol{v}^{\mathrm{F}}$ (appearance), the middle column shows $\mathbf{m}$ (shape) and the right column shows the joint sample where shape and appearance have been combined (red indicates the invisible part of the sample). *(c)* Test images with inferred masks $\mathbf{m}$ superimposed in semi-transparent red. The model largely identifies the foreground objects correctly, but struggles sometimes, especially if the background is poorly explained under the background model.

for the foreground appearance by training a Beta RBM directly on the full face images (i.e. including background; this pre-training does not teach the model to distinguish between foreground and background in the training images). For highly structured images the background models were sometimes not sufficiently powerful so that part of the background was assigned to the foreground even after consolidation of the foreground model. This does not completely prohibit learning of the foreground model but leads to a noisy final model. We addressed this issue by introducing an outlier component into the background model[6], i.e. each background pixel was modeled either by the background RBM or by a uniform distribution ($p = 0.3$), which can be incorporated into the Gibbs sampling scheme by modifying equations (3-5). Additional details can be found in the supplemental material available on the first author's homepage [13].

4 Results

Toy Data: For the toy data the model successfully learned about the shapes and appearances of the foreground objects: after learning, samples closely matched

---

[6] Using an "outlier-indicator" $o_i \in \{0, 1\}$ ($P(o_i = 1) = p$) the constraints in eq. (1) are replaced by $\left(\prod_i \delta \left[v_i^{\mathrm{F}} = x_i \right]^{m_i} \delta \left[v_i^{\mathrm{B}} = x_i \right]^{(1-m_i)(1-o_i)} \left[U(x_i) \right]^{(1-m_i)o_i} \right)$. For the toy dataset we initially trained only with the basic model and introduced the outlier component only for fine-tuning once the model had largely converged. For the faces we trained with the outlier component from the beginning.

Fig. 2. *(a)* Examples of the training data. *(b)*: Samples from the learned model. For the first three columns the format is similar to Fig. 1b, they demonstrate how shape ($\mathbf{m}$, left) and appearance ($\boldsymbol{v}^{\mathrm{F}}$, middle) combine to the joint sample (right). The remaining columns show further samples from the model. For the joint samples the red area is not part of the object. *(c)*: Inferred masks $\mathbf{m}$ (foreground-background segmentations) for a subset of the test images. Masks are superimposed on the test images in red. In most cases the model has largely correctly identified the pixels belonging to the face. Test images for which the model tends to make mistakes typically show the head in extreme poses. Labeling of the neck and the shoulders is somewhat inconsistent, which is expected given that there is considerable variability in the training images and that the model has not been trained to either include or exclude such areas. The same applies if parts of a face are occluded, e.g. by a hat. *(d)* Test images with random masks superimposed. If masks are randomly assigned to test images the alignment of mask and image is considerably worse. Additional training images and segmentation results and a comparison with results for a conventional conditional random field can be found in the supplemental material [13]. *(e)*: Convergence of the segmentation *during learning* (inferred mask $\mathbf{m}$ superimposed on training image before joint training (left most) and after 10, 20, 100 and 1000 epochs of joint training). At the beginning the segmentation is driven primarily by the background model and thus very noisy. *(f)* Two examples of the pairs of images used for the recognition task.

the foreground objects in the training set (cf. Fig. 1a) and inference on test patches led to largely correct segmentations of these patches into foreground and background: it labeled 96% of the pixels in our set of 5000 test images correctly [7] (see Fig. 1c for examples). To investigate to what extent the ability to ignore the background may help in recognition tasks we trained a simple RBM on the same data (same total number of hidden units as for the FG-BG model) and performed inference for a set of test patches in both models[8]. We trained a simple logistic classifier (with L2-regularization) on the inferred activation of the hidden units (only h^{F} for the FG-BG model) in order to classify patches whether they contain rectangles or round shapes. This an easy classification task if enough training data is available (e.g. 100 training patches per class), but as the number of training data points is reduced the classification performance drops strongly for the simple RBM but to a far lesser extent for the FG-BG model (classification performance 66% vs. 88% at 10 data points per class), suggesting that being able to ignore the background can help to improve recognition performance.

Faces: The model also learned a good representation of faces. Fig. 2b shows samples from the trained model. Although the samples do not exhibit as much detail as the faces in the training data (this is to be expected given the relatively small number of hidden units used) they exhibit important features of the training data, for instance, there are male and female faces, and the model has learned about different head positions and hair styles. Figure 2c shows segmentation results for a subset of the test images. In most cases the model has largely correctly identified the pixels belonging to the face. Test images for which the model tends to make mistakes typically show heads in extreme poses. Fig. 2d demonstrates that the model does not simply choose the same region in all images: randomly re-assigning the inferred masks to test images leads to considerably worse results. To investigate the effect of the mask in a very simple recognition task we manually segmented a small subset of the images in the dataset ($N = 65$). For each segmented person we created two test images by pasting the person into two different natural image background patches, thus obtaining two sets of images containing the same 65 different faces but against different backgrounds (two example pairs are shown in Fig. 2f). We inferred the segmentation and subsequently the hidden activation of the latent units of the foreground model. For each image from the first set we determined the most similar image in the second set (in terms of the RMS difference between the hidden unit activations). For 52 (80%) of the images in the first set the corresponding image with the same face (different background) in the second set was the closest match. This compares to 15 out of 65 (23%) for a simple Beta RBM (same number of hidden units as the foreground model) trained on the same dataset, suggesting that here, too, "ignoring" the background can lead to a representation in the hidden units that is less affected by the background than for a normal RBM.

---

[7] Chance is 50%; a CRF using a histogram of the background and contrast dependent smoothness term achieves 79%; see supplemental material [13] for further details.

[8] Inference in the simple RBM involves computing the activation of the hidden units given the test image; in the FG-BG model it involves inferring $\mathbf{m}$, $\boldsymbol{v}^{\mathrm{F}}$, and $\boldsymbol{v}^{\mathrm{B}}$.

5 Discussion

We have demonstrated how RBMs and layered representations can be combined to obtain a model that is able to represent the joint shape and appearance of foreground objects independently of the background and we have shown how to learn the model of the foreground directly from cluttered training images using only very weak supervision. The architecture is very flexible: it can be applied to images of different types (e.g. binary), the background model can be re-used for different foreground models, and it is possible to replace the background model independently of the foreground (e.g. if the statistics of the background change). Also, DBNs or DBMs could be used instead of RBMs. One interesting extension would be to include a third layer that is *in front* of the object layer. This would allow modeling occlusion of the foreground object (such occlusions, although rare in the face dataset, may explain part of the uncertainty in the learned shape model). Using *semi*-supervised schemes (e.g. with a few pre-segmented images) to learn more challenging object categories, is another exciting direction.

References

1. Hinton, G.E.: Training Products of Experts by Minimizing Contrastive Divergence. Neural Computation 14(8) 1771-1800 (2002)
2. Hinton, G.E., Osindero, S., Teh, Y.: A fast learning algorithm for deep belief nets. Neural Computation, 18, 1527-1554 (2006)
3. Lee, H., Gross, R., Ranganath, R., Ng, A.Y.: Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. ICML (2009).
4. Le Roux, N., Heess, N., Shotton J., Winn, J.: Learning a Generative Model of Images by Factoring Appearance and Shape. Neural Computation 23(3): 593-650 (2011)
5. Huang, G.B., Rames, M., Berg, T., Learned-Miller, E.: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. TR 07-49; Univ. of Mass., Amherst (2007)
6. Ranzato, M., Hinton, G.: Modeling Pixel Means and Covariances Using Factorized Third-Order Boltzmann Machines. CVPR (2010)
7. Salakhutdinov, R., Hinton, G. E.: Deep Boltzmann Machines. AISTATS (2009)
8. Tang, Y.: Gated Boltzmann Machine for Recognition under Occlusion. NIPS Workshop on Transfer Learning by Learning Rich Generative Models. (2010)
9. Tieleman, T.: Training Restricted Boltzmann Machines using Approximations to the Likelihood Gradient. ICML (2008)
10. Wang, J.Y.A., Adelson, E.H.: Representing moving images with layers. IEEE Transactions on Image Processing, 3(5):625 (1994)
11. Williams,C.K.I., Titsias,M.K.: Greedy Learning of Multiple Objects in Images using Robust Statistics and Factorial Learning. Neural Comp. 16(5) 1039-1062 (2004)
12. Wolf, L., Hassner, T., Taigman, Y.: Similarity Scores based on Background Samples. Asian Conference on Computer Vision (ACCV), Xi' an (2009)
13. Suppl. Material: http://homepages.inf.ed.ac.uk/s0677090/papers/icannSuppl.pdf