Bayesian Probabilistic Models for Image Retrieval

Vassilios Stathopoulos[1] Joemon M. Jose[2]

[1]Department of Statistical Science
University College London

[2]School of Computing Science
University of Glasgow

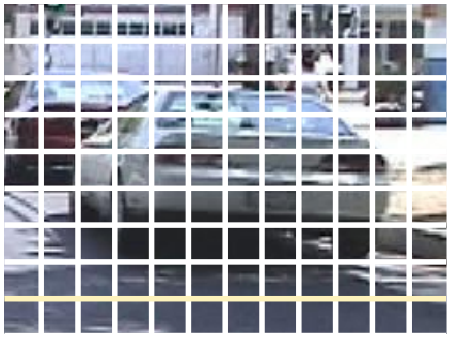
19-21 October 2011

Outline

Bag of Terms Image Retrieval

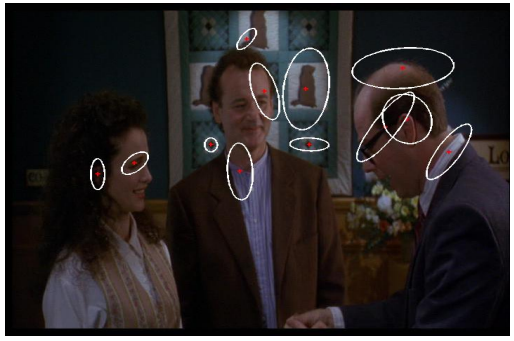
- Generate a representation that is similar to text documents.
- Images are represented by frequencies of parts.
- IR weighting and ranking functions can be directly applied to Bag of Terms models.
- Bag of Terms model relies on 3 stages
  1. Region Detection.
  2. Feature Description.
  3. Code-block Generation & Quantisation.

Region Detection



- Regular Grid

[Nowak et al.2006] [Tuytelaars2010]



- Interest Points

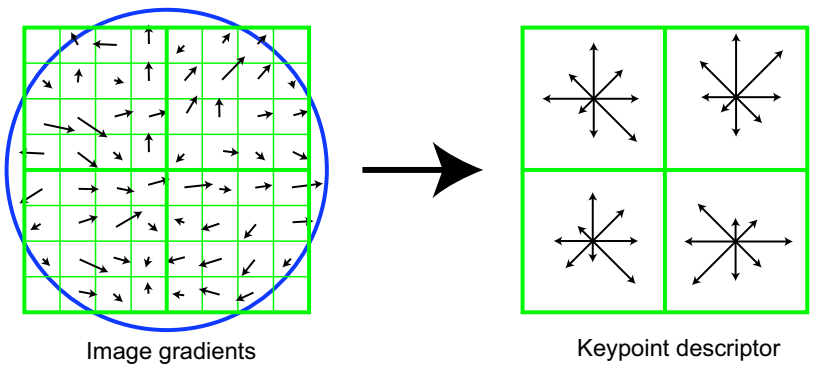
[Mikolajczyk et al.2005] [Csurka et al.2004]

- Segmentation

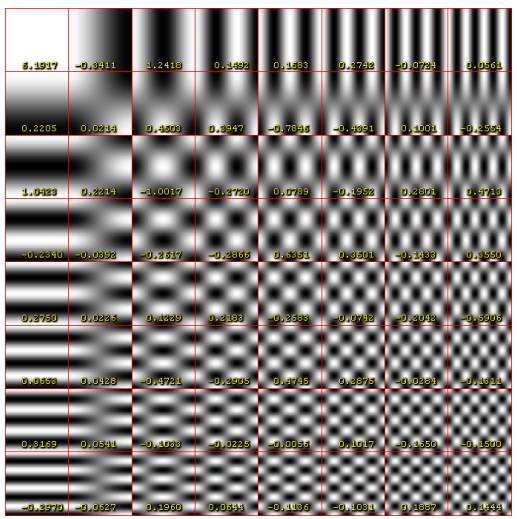
[Koniusz & Mikolajczyk2010]

Feature Description

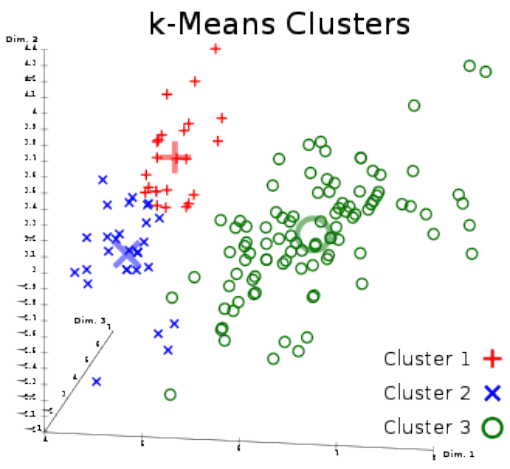
- Scale Invariant Feature Transform (SIFT) [Lowe2004]

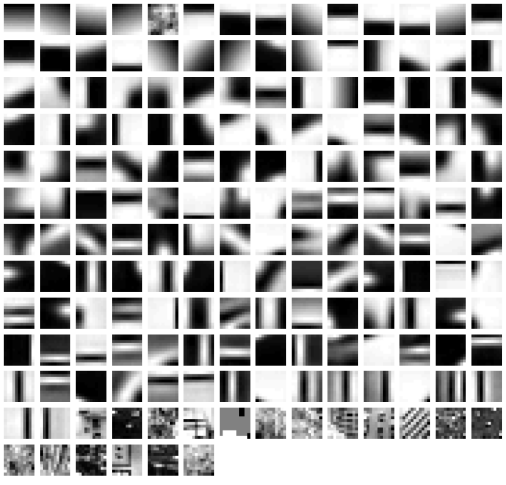


- Discrete Cosine Transform (DCT) [Carneiro et al.2007]



Code-block Generation & Quantisation





- Apply K-means to region feature descriptors from all collection images.
- Cluster means are treated as "*visual terms*".
- Quantise each image by associating each feature descriptor to its closest ""visual term".
- Images are represented as vectors $\mathbf{d} = \{d_1, \ldots, d_T\}$ where d_t is a "weight" of the importance of the t^{th} visual term.
- TF-IDF weighting $d_t = n_{t,d} \log \frac{N}{df_t}$

$$\text{score}(\mathbf{d}, \mathbf{q}) = \sum_t d_t \times q_t$$

Probabilistic IR Models

- Formal methodology for developing IR weighting and ranking algorithms.
- Rank documents / images based on the probability of relevance w.r.t. a user query.
- Two popular frameworks:
  1. Probabilistic Relevance Framework [Robertson & Zaragoza2009]
  2. Language Modeling Framework [Hiemstra2001]

Language Models for IR

- Assume a generative process for each document in the collection.

$$p(\mathbf{d}|\boldsymbol{\theta}_d) = \mathcal{M}(\mathbf{d}|\boldsymbol{\theta}_d) = \frac{(\sum_t n_{t,d})!}{\prod_t n_{t,d}!} \prod_t \theta_{d,t}^{n_{d,t}}$$

- ML estimate for $\theta_{d,t} = n_{d,t} / \sum_{t'} n_{d,t'}$ leads to over-fitting problems for terms with 0 frequency.
- Introduce a Dirichlet prior $\mathcal{D}(\boldsymbol{\theta}_d|\boldsymbol{\alpha})$ over model parameters and obtain a MAP estimate

$$\hat{\boldsymbol{\theta}}_d^{(MAP)} = \underset{\boldsymbol{\theta}_d}{\text{argmax}}\, p(\mathbf{d}|\boldsymbol{\theta}_d) p(\boldsymbol{\theta}_d), \quad \hat{\theta}_{d,t}^{(MAP)} = \frac{(n_{d,t} + \alpha_t - 1)}{\sum_{t'} (n_{d,t'} + \alpha_{t'} - 1)}$$

- Prior parameters α_t are usually set to the average frequency of the t^{th} term in the collection.

Language Models for IR

- Give a query $\mathbf{q}$ rank documents using the query likelihood

$$\log p(\mathbf{q}|\hat{\theta}_d^{(MAP)}) \quad \propto_q \quad \sum_{\{t:n_{q,t}>0 \wedge n_{d,t}>0\}} n_{q,t} \log\left(\frac{n_{d,t}}{\alpha_t - 1} + 1\right)$$
$$- \quad \log\left(\sum_{t'} n_{d,t'} + \alpha_{t'} - 1\right) \sum_{\{t:n_{q,t}>0\}} n_{q,t}$$

- Ranking function depends only on terms common in the document and query.
- Efficient implementation with an inverted index data structure.

Probabilistic Models for Image Retrieval

- Model the density of continuous image features directly using semi-parametric models.
- Images are unordered sets of vectors
  $\mathbf{d} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{N_I}\}, \quad \mathbf{x} \in \mathbb{R}^D$
- Gaussian Mixture Models,
  [Westerveld et al.2003, Vasconcelos & Lippman et al.2003]

$$p(\mathbf{d}|\boldsymbol{\theta}_d) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

- Maximum likelihood parameter estimates using the EM algorithm.
- Given a query image $\mathbf{q}$ rank images using the query likelihood $\log p(\mathbf{q}|\hat{\theta}_d^{(ML)})$.
- No efficient data structure

Model Predictive Density

$$p(\mathbf{x}^*|\mathbf{d}) = \int p(\mathbf{x}^*|\boldsymbol{\theta}) \underbrace{p(\boldsymbol{\theta}|\mathbf{d})}_{\text{posterior}} d\boldsymbol{\theta}$$

- Marginalise uncertainty about the parameters $\boldsymbol{\theta}$.
- MAP and ML estimates can be seen as approximations of the predictive density.

$$p(\mathbf{x}^*|\mathbf{d}) \approx p(\mathbf{x}^*|\hat{\boldsymbol{\theta}}_d^{(MAP)})$$

- Point estimates are asymptotically $n \to \infty$ optimal.
- Images and documents only contain a finite set of observations.
- Number of parameters is usually large, e.g. in the order of vocabulary terms.

Multinomial-Dirichlet Model

- The posterior for the Multinomial-Dirichlet model is a Dirichelt

$$p(\boldsymbol{\theta}_d|\mathbf{d}) = \frac{p(\mathbf{d}|\boldsymbol{\theta}_d)p(\boldsymbol{\theta}_d)}{\int p(\mathbf{d}|\boldsymbol{\theta}_d)p(\boldsymbol{\theta}_d)d\boldsymbol{\theta}_d} = \mathcal{D}(\boldsymbol{\theta}_d|\mathbf{n}_{d,\cdot} + \boldsymbol{\alpha})$$

- The predictive density is also available in closed form [Zaragoza et al.2003] and its log is proportional to

$$\begin{aligned}\log p(\mathbf{q}|\mathbf{d}) \quad \propto_q & \sum_{t:n_{t,q}>0 \wedge n_{t,d}>0} \sum_{g=1}^{n_{t,q}} \log\left(\frac{n_{t,d}}{\alpha_t + g - 1} + 1\right) \\ - & \sum_{j=1}^{\sum_{t'} n_{t',q}} \log\left(\sum_{t'} n_{t',d} + \alpha_{t'} + j - 1\right)\end{aligned}$$

- Ranking function depends only on terms common in the document and query.

Gaussian Mixture Model

- Posterior is not tractable for mixture models. Two possible approaches:
  1. MCMC samples from the posterior.
  2. Variational approximation.
- MCMC is asymptotically optimal as the number of samples tends to infinity.
- Several chains have to run for each image in the collection to monitor convergence.
- For a query the predictive density is the weighted sum of the posterior samples.
- Variational approach provides a *"local"* approximation to the posterior.
- Posterior and predictive density have convenient analytical forms.

Variational Inference for Gaussian Mixture Model

- Latent variable representation

$$p(\mathbf{d}|\boldsymbol{\theta}_d, \mathbf{Z}) = \prod_{n=1}^{N} \prod_{k=1}^{K} [\pi_k \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)]^{z_{n,k}}$$

- Conjugate prior
- $p(\boldsymbol{\pi}) = \mathcal{D}(\boldsymbol{\pi}|\alpha_0)$, small α_0 gives preference to "sparse" solutions.
- $p(\boldsymbol{\mu}_k|\boldsymbol{\Sigma}_k) = \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_0, \beta^{-1}\boldsymbol{\Sigma}_k)$. $\mathbf{m}_0$ can be set to the mean of feature descriptors in the collection. A small β ensures prior is flat in high likelihood regions.
- $p(\boldsymbol{\Sigma}_k) = \mathcal{IW}(\boldsymbol{\Sigma}_k|\mathbf{W}_0, v_0)$. $\mathbf{W}_0$ can be set to the precision of feature descriptors in the collection. Set v_0 such that prior is flat in high likelihood regions.

Variational Inference for Gaussian Mixture Model

- Augment parameters and latent variables $\mathbf{\Theta} = \{\boldsymbol{\theta}_d, \mathbf{Z}\}$
- Consider an approximate posterior that factorizes such that $q(\mathbf{\Theta}) = q(\boldsymbol{\theta}_d)p(\mathbf{Z})$
- Applying Jensen's inequality the marginal can be written

$$p(\mathbf{d}) = \underbrace{\int q(\mathbf{\Theta}) \log \frac{p(\mathbf{d}, \mathbf{\Theta})}{q(\mathbf{\Theta})} d\mathbf{\Theta}}_{\text{Lower Bound}} - \underbrace{\int q(\mathbf{\Theta}) \log \frac{p(\mathbf{\Theta}|\mathbf{d})}{q(\mathbf{\Theta})} d\mathbf{\Theta}}_{\text{KL}}$$

- By maximising the *Lower Bound* the KL is minimised.
- $q(\mathbf{\Theta}_d)$ can be further factored as $q(\mathbf{Z})q(\boldsymbol{\pi}) \prod_{k=1}^{K} q(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$.
- Taking each factor separately while considering all others constant we can iteratively optimise the lower bound, e.g.

$$\log q(\mathbf{Z}) = \int \log p(\mathbf{d}, \mathbf{\Theta}) q(\boldsymbol{\theta}_d) d\boldsymbol{\theta}_d + const$$

Variational Inference for Gaussian Mixture Model

- The variational posterior takes the following form [Bishop2006, Chap. 7]

$$\begin{aligned} q(\mathbf{z}_n) &= \mathcal{M}(\mathbf{z}_n|1,\mathbf{r}_n) \\ q(\boldsymbol{\pi}) &= \mathcal{D}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \\ q(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k) &= \mathcal{N}(\boldsymbol{\mu}_k|\mathbf{m}_k,\beta_k^{-1}\boldsymbol{\Sigma}_k)\mathcal{IW}(\boldsymbol{\Sigma}_k|\mathbf{W}_k,v_k) \end{aligned}$$

- The parameters of the variational posterior $\boldsymbol{\alpha},\boldsymbol{\rho},\mathbf{m},\mathbf{W}$ are optimised using the Variational EM algorithm (VEM) [Bishop2006, Chap. 7].
- The predictive density can also obtained explicitly

$$\begin{aligned} p(\mathbf{x}^*|\mathbf{d}) &= \sum_{k=1}^{K}\int\int\int \pi_k\mathcal{N}(\mathbf{x}^*|\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)q(\boldsymbol{\pi})q(\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)d\boldsymbol{\pi}d\boldsymbol{\mu}_k d\boldsymbol{\Sigma}_k \\ &= \frac{1}{\hat{\boldsymbol{\alpha}}}\sum_{k=1}^{K}\alpha_k \mathrm{St}\left(\mathbf{x}^*|\mathbf{m}_k,\frac{(v_k+1-D)\beta_k}{1+\beta_k}\mathbf{W}_k,v_k+1-D\right) \end{aligned}$$

Determining the Number of Components

- From the Dirichlet variational posterior over the mixing coefficients $\boldsymbol{\pi}$ we have

$$\mathbb{E}[\pi_k] = \frac{\alpha_k}{\hat{\boldsymbol{\alpha}}}, \quad \text{Var}(\pi_k) = \frac{\alpha_k(\hat{\boldsymbol{\alpha}} - \alpha_k)}{\hat{\boldsymbol{\alpha}}^2(\hat{\boldsymbol{\alpha}} + 1)}, \quad \hat{\boldsymbol{\alpha}} = \sum_{k=1}^{K} \alpha_k$$

- In the VEM algorithm the α_k parameters are updated as

$$\alpha_k = \alpha_0 + \sum_{n=1}^{N} r_{n,k}$$

- When a_0 is small, set K to a relatively large value and remove components with $\alpha_k = \alpha_0$ [Bishop & Corduneanu 2001] as they have negligible contribution to the predictive density.

Corel 5K Test Collection

- 4,500 training images, 500 test images.
- Collection is divided into 50 categories, e.g. “sunset”, “roses”, “stamps” etc.
- We index the 4,500 training images which contain 90 images per category.
- The 500 test images are used as queries, 10 images for each category.
- Given a query image we expect the 90 images from corresponding category to be ranked first.

Pre-processing

- Images are converted to the YUV colour space. 1 Luminance and 2 chrominance channels.
- Segment images using a 8×8 pixels sliding window with 4 pixels overlap.
- DCT is applied to each 8×8 pixels region.
- For the Bag of Terms representation we used K-means with 2,000 clusters.
- For the GMM the EM algorithm was used with 8 components [Westerveld et al.2003].
- For the VEM the number of components was initially set to 40 and then components were removed.
- The EM and VEM algorithms where initialised by randomly setting the latent variables $\mathbf{Z}$.

Results

Table: Retrieval results for 500 query images in the test set. $*$ indicates statistical significance using a Wilcoxon rank-sum test with 1% significance level.

Method	**MAP**	**R-Prec.**	**P@5**	**P@10**	**P@20**
BOT-MAP	0.0333	0.0364	0.0441	0.0429	0.0383
BOT-PD	0.0341	0.0375	0.0477	0.0431	0.0387
GMM-ML	**0.0975***	**0.1280***	**0.3038***	**0.2599***	**0.2179***
GMM-MAP	0.0999	0.1308	0.3070	0.2645	0.2210
GMM-PD	**0.1165***	**0.1457***	**0.3315***	**0.2836***	**0.2370***

Results

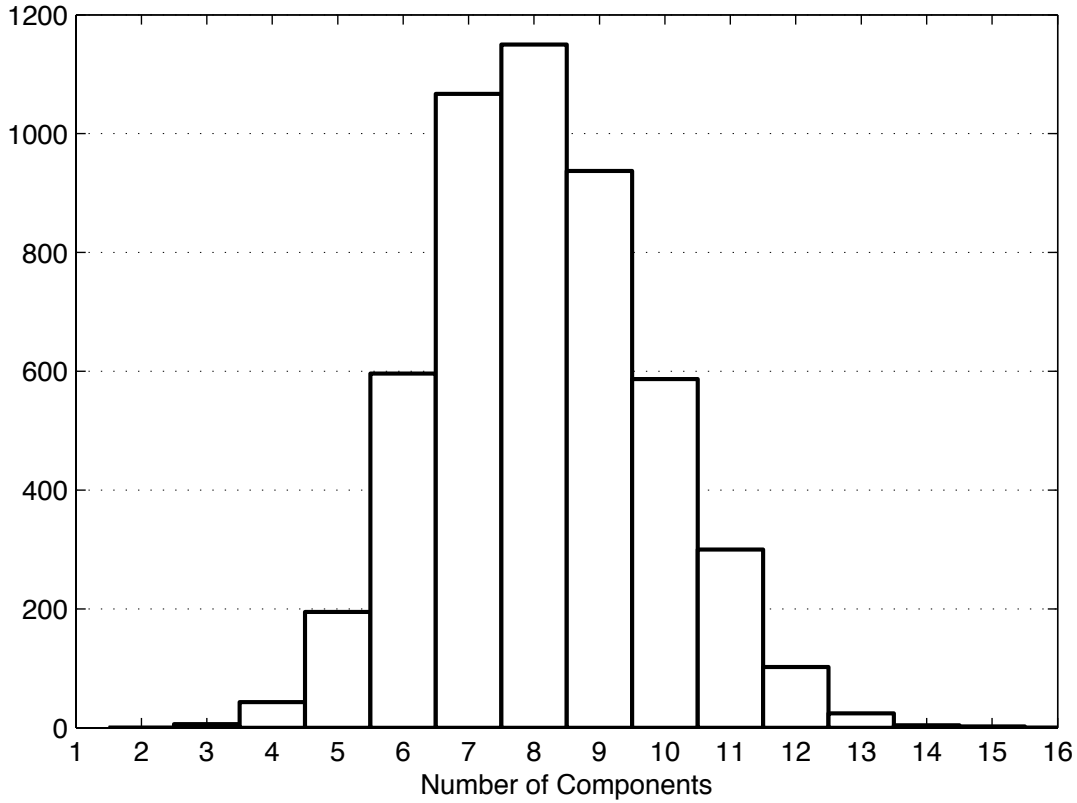


Figure: Distribution of the number of components K for the 4.500 images in the collection.

Conclusions

- Scalability of the Bag of Terms representation is questionable as quantisation of query images is required.
- K-means code-block generation is computationally challenging. Alternatives, DBSCAN, hierarchical clustering.
- Quantisation errors can significantly decrease retrieval effectiveness.
- Probabilistic image retrieval models are superior to Bag of Terms approaches.
- Retrieval requires a linear scan through the collection.
- The predictive density ranking function is always superior w.r.t. ML and MAP estimates, indicative of over-fitting.
- Number of mixture components can be identified automatically from the data.
- VEM has the same order of complexity as the EM algorithm.

Future Work

- Improve indexing structure for probabilistic retrieval models.
- Locality Sensitive Hashing (LSH) on Kernel spaces [Kulis & Grauman 2009].
- Sub linear complexity with theoretical approximation error bounds.
- Kernel functions for probabilistic generative models.
- Fisher Kernels [Jaakkola & Haussler1999], Probability Product Kernels [Jebara et al.2004]

References

Eric Nowak and Frédéric Jurie and Bill Triggs

Sampling strategies for bag-of-features image classification.

ECCV, 2006.

Tuytelaars, T.

Dense interest points

CVPR, 2010.

Krystian Mikolajczyk and Tinne Tuytelaars and Cordelia Schmid and Andrew Zisserman and Jiri Matas and Frederik Schaffalitzky and Timor Kadir and Luc J. Van Gool

A Comparison of Affine Region Detectors

IJCV, 2005.

Gabriela Csurka and Christopher R. Dance and Lixin Fan and Jutta Willamowski and Cédric Bray

Visual categorization with bags of keypoints

ECCV, 2004.

Koniusz, Piotr and Mikolajczyk, Krystian

On a Quest for Image Descriptors Based on Unsupervised Segmentation Maps

ICPR, 2010.

David G. Lowe

Distinctive Image Features from Scale-Invariant Keypoints

IJCV, 2004.

Gustavo Carneiro and Antoni B. Chan and Pedro J. Moreno and Nuno Vasconcelos

Supervised Learning of Semantic Classes for Image Annotation and Retrieval

TPAMI, 2007.

Thijs Westerveld and Arjen P. de Vries and Alex van Ballegooij and Franciska de Jong and Djoerd Hiemstra

A probabilistic multimedia retrieval model and its evaluation

EURASIP: JASP, 2003.

Nuno Vasconcelos and Andrew Lippman

A Probabilistic Architecture for Content-Based Image Retrieval

CVPR, 2000.

References

Robertson, Stephen and Zaragoza, Hugo

The Probabilistic Relevance Framework: BM25 and Beyond

Found. Trends Inf. Retr., (3) 2009.

Djoerd Hiemstra

Using Language Models for Information Retrieval

PhD Thesis, University of Twente, 2001.

Zaragoza, Hugo and Hiemstra, Djoerd and Tipping, Michael

Bayesian language model, ad hoc language model, ad hoc retrieval, information retrieval

SIGIR, 2003.

Bishop, Christopher M.

Pattern Recognition and Machine Learning

Springer, 2006.

Bishop, C. M. and Corduneanu, A.

Variational Bayesian model selection for mixture distributions

Artificial Intelligence and Statistics, 2006.

Brian Kulis and Kristen Grauman

Kernelized locality-sensitive hashing for scalable image search

ICCV, 2009.

Jebara, Tony and Kondor, Risi and Howard, Andrew

Probability Product Kernels

JMLR, 2004.

Jaakkola, Tommi S. and Haussler, David

Exploiting generative models in discriminative classifiers

NIPS, 1999.