Soft Assignment Of Visual Words As Linear Coordinate Coding And Optimisation Of Its Reconstruction Error

{P.Koniusz, K.Mikolajczyk}@surrey.ac.uk

UNIVERSITY OF SURREY

BBC R&D

Aim

- **Visual Word Uncertainty** [1] (also referred to as **Soft Assignment** and **Kernel Codebook**) is a well established technique for representing images as histograms by flexible assignment of image descriptors to a visual vocabulary
- This work investigates a connection between Visual Word Uncertainty and more recent Linear Coordinate Coding [2] methods
- As a result, parameters of Kernel Codebook are learnt by minimising corresponding feature coding error
- This yields state-of-the-art results with bags-of-words on VOC 2010 Action Recognition dataset [4] and in the object category classification

Motivation

- Better understanding of Soft Assignment in context of Gaussian Mixture Models [3]
- Desire to improve quantisaton properties of Soft Assignment coding
- Finding connection between the parameter space of Soft Assignment and Linear Coordinate Coding
- Need for a robust way of minimising associated Quantisation Error (Reconstruction Error)

Key Ideas

- Viewing Soft Assignment as a simplified Gaussian Mixture Model
- Benefiting from Membership Probabilities as approximation to Linear Coding
- Quantisation Error is synonymous with Residual Error between encoded features and resulting codes being back-projected into feature space
- Recognising the posed problem as convex in its nature

References

[1] J. C. Van Gemert, C. J. Veenman, A. W. M. Smeulders, J. M. Geusebroek. Visual Word Ambiguity. In *Pattern Analysis and Machine Intelligence*

[2] K. Yu, T. Zhang, Y. Gong. Nonlinear Learning using Local Coordinate Coding. In *Neural Information Processing Systems*

[3] J. Bilmes. A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. *Technical Report*

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. *http://pascallin.ecs.soton.ac.uk/challenges/VOC*

[5] M. E. Nilsback, A. Zisserman. Automated Flower Classification over a Large Number of Classes. In *International Conference on Computer Vision*

[6] P. Koniusz, K. Mikolajczyk. On a Quest for Image Descriptors Based on Unsupervised Segmentation Maps. In *International Conference on Pattern Recognition*

Model

- Gaussian Mixture Model and corresponding Component Membership Probabilities can be expressed as follows:

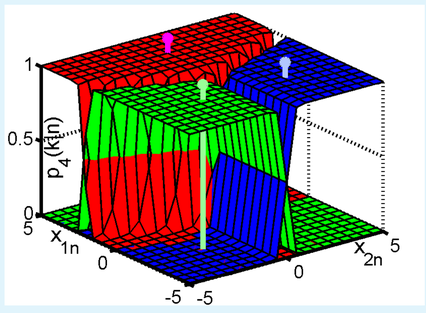
$$\Lambda(X;\theta)=\prod_{n=1}^{N}\sum_{k=1}^{K}p_k g(\boldsymbol{x}_n;\boldsymbol{m}_k,\boldsymbol{\sigma}_k) \quad \Rightarrow \quad p(k|n)=\frac{p_k g(\boldsymbol{x}_n;\boldsymbol{m}_k,\boldsymbol{\sigma}_k)}{\sum_{k'=1}^{K}p_{k'} g(\boldsymbol{x}_n;\boldsymbol{m}_{k'},\boldsymbol{\sigma}_{k'})} \tag{1}$$

- K denotes the number of Gaussian components, $p_{k\in\{1,..,K\}}$ are the component mixing probabilities, $\boldsymbol{m}_k$ are the Gaussian means, $\boldsymbol{\sigma}_k$ are the component standard deviations, and $\boldsymbol{x}_{n\in\{1,..,N\}}$ are N descriptors of a dataset
- The parameters $\theta=(\theta_1,...,\theta_K)=((p_1,\boldsymbol{m}_1,\boldsymbol{\sigma}_1),...,(p_K,\boldsymbol{m}_K,\boldsymbol{\sigma}_K))$ of the model consist of a vast number of degrees of freedom and therefore can be further reduced to $\theta=(\theta_1,...,\theta_K)=((\boldsymbol{m}_1,\sigma),...,(\boldsymbol{m}_K,\sigma))$ by fixing all mixing probabilities $p_1=p_2=...=p_K\neq 0$ and $\boldsymbol{\sigma}_1=\boldsymbol{\sigma}_2=...=\boldsymbol{\sigma}_K=\sigma\neq 0$. This yields the membership probabilities as follows:

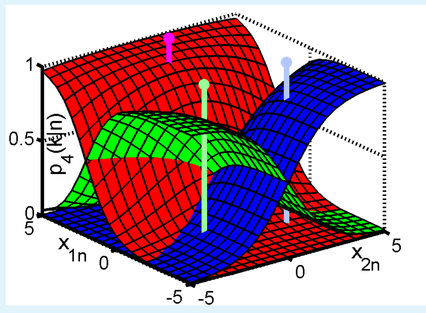
$$p(k|n)=\frac{g(\boldsymbol{x}_n;\boldsymbol{m}_k,\sigma)}{\sum_{k'=1}^{K}g(\boldsymbol{x}_n;\boldsymbol{m}_{k'},\sigma)} \tag{2}$$

- The approximation of a vector $\boldsymbol{x}$ can be expressed as $\widetilde{\boldsymbol{x}}=\sum_{\boldsymbol{m}\in M}\gamma_{\boldsymbol{m}}(\boldsymbol{x})\boldsymbol{m}$
- The residual error of approximation of a descriptor $\boldsymbol{x}_n$ is $\xi_n^2=\left\|\boldsymbol{x}_n-\sum_{\boldsymbol{m}\in M}\gamma_{\boldsymbol{m}}(\boldsymbol{x}_n)\boldsymbol{m}\right\|^2$
- Using linear codes $\gamma_{\boldsymbol{m}_k}(\boldsymbol{x}_n)=p(k|n)$ we can gauge Quantisation Error of Soft Assignment
- Therefore, total ξ^2 can be minimised with respect to Soft Assignment Smoothing Factor σ as follows:

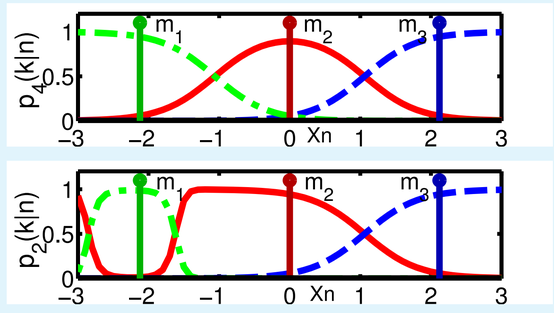
$$\min_{\sigma}\sum_{n=1}^{N}\left\|\boldsymbol{x}_n-\sum_{k=1}^{K}\frac{g(\boldsymbol{x}_n;\boldsymbol{m}_k,\sigma)}{\sum_{k'=1}^{K}g(\boldsymbol{x}_n;\boldsymbol{m}_{k'},\sigma)}\boldsymbol{m}_k\right\|^2 \tag{3}$$



$p(k|n)$ for $\sigma^2=1$ (simplified GMM, in 2D)



$p(k|n)$ for $\sigma^2=9$ (simplified GMM, in 2D)

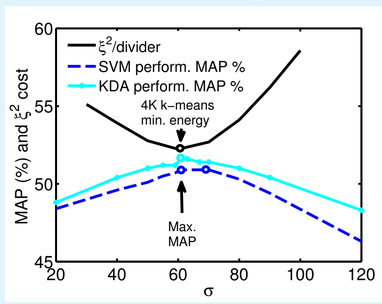


(**top**) simplified GMM (**bottom**) full GMM

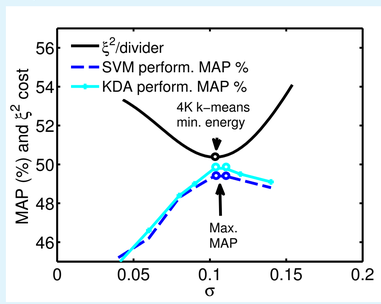
- Empirically, one can see $p(k|n)$ has close to linear responses in the above plots
- Full GMM (**right bottom plot**) produces non-unique codes that deteriorate reconstruction

Results and Conclusions

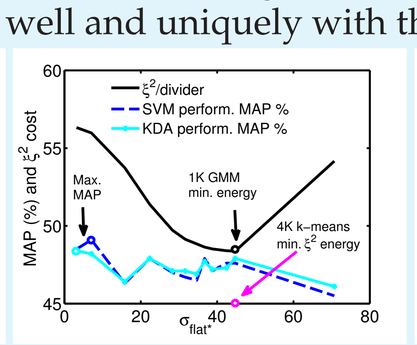
We evaluate the proposed σ-optimisation scheme on VOC 2010 Action Recognition [4] and Flower17 [5] datasets. Results are reported as a function of σ using Mean Average Precision. Reconstruction Error ξ^2 (in black) is shown to correlate well and uniquely with the MAP peaks.



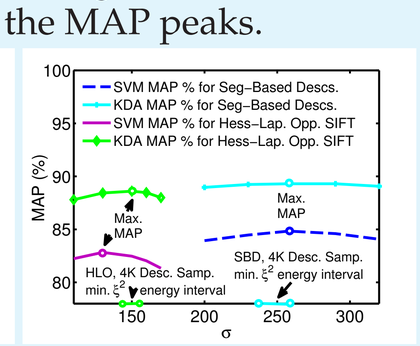
MAP maxima and ξ^2 minima (VOC2010, k-means, two variants of SIFT, Soft Assignment eq. 2)



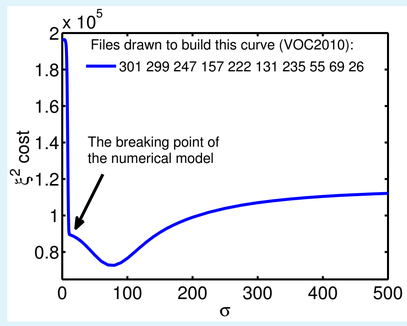
MAP maxima and ξ^2 for GMM given by equation 1



MAP maxima and ξ^2 minima intervals on Flower17 (Opp. SIFT, Seg-Based Desc. [6])



- The highest MAP performance is attained at minima of ξ^2 for the simplified GMM given in equation 2
- Full GMM in equation 1 results in many local MAP maxima
- The associated reconstruction cost remains a convex problem
- Optimisation in equation 3 can be performed efficiently with just a small subset of all descriptors by the gradient descent
- Some datasets may further thrive on discriminative σ learning



Cost as a function of σ (note its convexity)

Acknowledgements

This work was sponsored by the BBC Future Media and Technology and EPSRC EP/F003420/1 research grants.