SOFT ASSIGNMENT OF VISUAL WORDS AS LINEAR COORDINATE CODING AND OPTIMISATION OF ITS RECONSTRUCTION ERROR

Piotr Koniusz and Krystian Mikolajczyk

University of Surrey, Guildford, UK

ABSTRACT

Visual Word Uncertainty also referred to as Soft Assignment is a well established technique for representing images as histograms by flexible assignment of image descriptors to a visual vocabulary. Recently, an attention of the community dealing with the object category recognition has been drawn to Linear Coordinate Coding methods. In this work, we focus on Soft Assignment as it yields good results amidst competitive methods. We show that one can take two views on Soft Assignment: an approach derived from Gaussian Mixture Model or special case of Linear Coordinate Coding. The latter view helps us propose how to optimise smoothing factor of Soft Assignment in a way that minimises descriptor reconstruction error and maximises classification performance. In turns, this renders tedious cross-validation towards establishing this parameter unnecessary and yields it a handy technique. We demonstrate state-of-the-art performance of such optimised assignment on two image datasets and several types of descriptors.

Index Terms— Image classification, soft assignment, coordinate coding, descriptor reconstruction error, bags-of-words

1. INTRODUCTION

Transforming local image descriptors into histograms lies at a heart of the object category recognition. The search for appropriate coding schemes expressing optimally content of images has been a subject of recent activity in the community. A number of methods have been proposed up-to-date including Hard and Soft Assignment [1], family of Linear Coordinate Coding [2] entailing Sparse Coding [3] and Local Coordinate Coding [4], and approaches like Fisher Kernels [5] or Super Vector Coding [6].

Hard Assignment associates each descriptor vector with the nearest visual word of a given dictionary. Whilst this provides with a reasonable expressive power, a single descriptor belongs to only one closest word in a dictionary. This yields a high quantisation error. Soft Assignment mitigates this effect by allowing soft contribution of each descriptor to its closest words in a dictionary. This was initially achieved by a crude heuristic like assigning a given descriptor to k-nearest words, all with equal weights. Subsequently, Visual Word Uncertainty [1] was found a more appropriate weighting scheme, though, with one inconvenient parameter to evaluate in n-fold classification on validation data, i.e. fivefold cross validation results in 5x more computations. To further reduce quantisation errors, Linear Coordinate Coding [2] was proposed. It expresses each descriptor vector as a sparse linear combination of neighbouring dictionary anchors. L1 regularisation norm over assignments favours only a small subset of activations leading to sparsity. Also, Sparse Coding [3] with Spatial Pyramid Match and the maximum pooling produced attractive results.

In this paper, we bridge the gap in understanding of Soft Assignment (SA) in the context of Linear Coordinate Coding (LCC) as the first approach can be viewed as a particular subset of solutions of the latter, provided no sparsity is forced. We also relate such weighting scheme to Component Membership Probabilities of Gaussian Mixture Models (GMM) [7]. Next, we exploit foundations of LCC to find an optimal smoothing factor for SA yielding lower quantisation errors. We show it leads to top classification results on two datasets. Also, we demonstrate that using GMM with multiple variances results in a higher reconstruction error suggesting full-parameter GMMs [7] are more suitable to work with other techniques, i.e. Fisher Kernels [5].

2. SOFT ASSIGNMENT AND RELATION TO GMMS.

For a Mixture of K Gaussian functions and mixing probabilities, one can express parameters of GMM [7] to estimate as $\theta = (\theta_1, ..., \theta_K) = ((p_1, \boldsymbol{m}_1, \boldsymbol{\sigma}_1), ..., (p_K, \boldsymbol{m}_K, \boldsymbol{\sigma}_K))$ and address density estimation problem by optimising:

$$\Lambda(X;\theta) = \prod_{n=1}^{N} \sum_{k=1}^{K} p_k g(\boldsymbol{x}_n; \boldsymbol{m}_k, \boldsymbol{\sigma}_k) \tag{1}$$

K denotes number of components, $p_{k\in\{1,..,K\}}$ are component mixing probabilities, $\boldsymbol{m}_k$ are Gaussian means, $\boldsymbol{\sigma}_k$ are component deviations, and $\boldsymbol{x}_{n\in\{1,..,N\}}$ are the descriptors of a dataset. In turns, the membership probability of component k being induced given descriptor $\boldsymbol{x}_n$ with index n is:

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

However, if parameters to estimate are $\theta = (\theta_1, ..., \theta_K) = ((\boldsymbol{m}_1, \sigma), ..., (\boldsymbol{m}_K, \sigma))$, the density estimation cost function can be rewritten as:

$$\Lambda(X; \theta) = \prod_{n=1}^{N} \sum_{k=1}^{K} g(\boldsymbol{x}_n; \boldsymbol{m}_k, \sigma) \tag{3}$$

Therefore, the membership probability equation 2 becomes:

$$p(k|n) = \frac{g(\boldsymbol{x}_n; \boldsymbol{m}_k, \sigma)}{\sum_{k'=1}^{K} g(\boldsymbol{x}_n; \boldsymbol{m}_{k'}, \sigma)} \tag{4}$$

This is a well-known expression for the Soft Assignment (SA) [1] that is used in forming histograms. For every $k \in K$, corresponding expected value of $p(k|n)$ over all $\boldsymbol{x}_n$ of a given image yields an entry to a k-th final histogram bin. One could assume that finding the optimal σ is a subject to optimising the cost in equation 3. We found that means $\boldsymbol{m}_k$ produced in such process tend to be of better quality than those estimated by k-means leading to better results. However, σ estimated in such way proved severely underestimated. This indirectly suggests that full-parameter GMM given in equation 1 would suffer from similar issue if histograms were to be built using GMM estimated $\boldsymbol{\sigma}_k$ applied to membership probabilities in equation 2. We show empirically this is the case in the experiment section.

3. SOFT ASSIGNMENT AND LINEAR COORDINATE CODING

The foundations of Linear Coordinate Coding are provided in [2]. We discuss only the formulations essential to our work. Coordinate Coding is a pair (γ, M), where $M \subset \mathbb{R}^D$ is a set of visual words of a given dictionary and γ is a mapping of descriptor vector $\boldsymbol{x} \in \mathbb{R}^D$ to a vocabulary vector $[\gamma_{\boldsymbol{m}}(\boldsymbol{x})]_{\boldsymbol{m} \in M} \in \mathbb{R}^{K=|M|}$. One can further impose $\sum_{\boldsymbol{m}} \gamma_{\boldsymbol{m}}(\boldsymbol{x}) = 1$ and $\gamma_{\boldsymbol{m}}(\boldsymbol{x}) \geq 0$ if this is to produce a histogram per a descriptor vector. The approximation of $\boldsymbol{x}$ 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 vector $\boldsymbol{x}_n$ can be expressed as:

$$\xi_n^2 = \|\boldsymbol{x}_n - \sum_{\boldsymbol{m} \in M} \gamma_{\boldsymbol{m}}(\boldsymbol{x}_n) \boldsymbol{m}\|^2 \tag{5}$$

The approximation error of all descriptors can be expressed as expected value of terms ξ_n^2 over all $\boldsymbol{x}_n$ or simply as a sum $\xi^2 = \sum_n \xi_n^2$. Such defined error is equivalent to the quantisation error. Therefore, plugging equation 4 into equation 5 yields a cost function we seek to minimise with respect to σ:

$$\min_{\sigma} \sum_{n=1}^{N} \left\| \boldsymbol{x}_n - \sum_{k=1}^{K} \frac{g(\boldsymbol{x}_n; \boldsymbol{m}_k, \sigma)}{\sum_{k'=1}^{K} g(\boldsymbol{x}_n; \boldsymbol{m}_{k'}, \sigma)} \boldsymbol{m}_k \right\|^2 \tag{6}$$

Generally, there is a strong relation between equation 6 and Linear Coordinate Coding [2]. The latter scheme is optimised with respect to assignment coefficients $\gamma_{\boldsymbol{m}}(\boldsymbol{x}_n)$ to achieve a good linear combination of anchors $\boldsymbol{m}$ approximating a given descriptor $\boldsymbol{x}_n$. Soft Assignment (SA) can be also viewed as approximating descriptors if linear combinations of anchors $\boldsymbol{m}$ weighted by assignment coefficients $p(k|n)$ are applied to evaluate the residual error to find the optimal σ. The assignments are taken from the space spanned by the membership probabilities given by equation 4. Note, membership probabilities in figure 1 (top and bottom left) have almost linear slopes (subject to well-chosen σ) and spanned locally for descriptors falling between neighbouring spanning anchors. This makes it somewhat similar to Local Coordinate Coding [4]. However, if full GMM membership probabilities (equation 2) are used as in figure 1 (bottom right), the locality property becomes violated (red solid and green dashed curves). Further, slopes become ill-spanned resulting in a poor approximation of descriptors in proximity of $\boldsymbol{m}_2$. The reconstruction emphasis is put on descriptors in proximity of the narrow peak despite these descriptors differ from each other by poor SNR. This is why SA with full GMM performs poorly. Update rule for σ based on equation 3 is closely related to equation 6, however, the differences suggest σ has two different meanings in case of i) the optimal reconstruction of descriptor vectors gauged by ξ^2 and ii) the density estimation. Practical optimisation of equation 6 is achieved by applying a generic optimiser with Gradient and Hessian of equation in question: $\frac{\partial \xi^2}{\partial \sigma} = [\xi^2(\sigma+\Delta\sigma) - \xi^2(\sigma-\Delta\sigma)]/2\Delta\sigma$, $\frac{\partial^2 \xi^2}{\partial \sigma^2} = [\xi^2(\sigma+\Delta\sigma) + \xi^2(\sigma-\Delta\sigma) - 2\xi^2(\sigma)]/(\Delta\sigma)^2$. Value of $\Delta\sigma$ depends on the descriptors used in the experiments outlined in section 4. It determines the quality of the gradient approximation and is set arbitrarily to 1 and 0.001 for large and unit norm descriptors. Similarly to GMMs [7], there is no closed form solution for equation 6. It suffices to mention at this stage that the cost function remains convex in σ and that only small subset of descriptor vectors from a dataset has to be evaluated. This is shown in the experimental section.

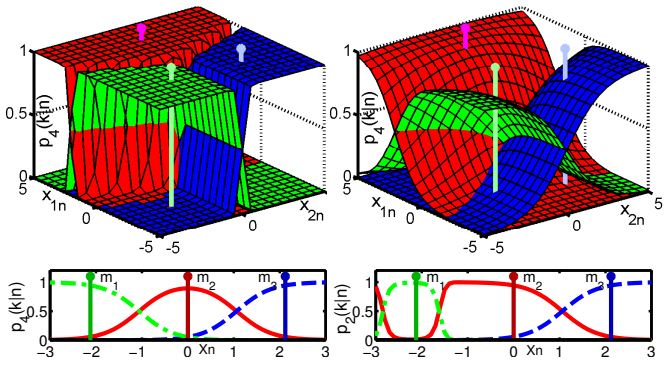

Fig. 1. (Top) Space of membership probabilities given by equation 4 for three arbitrarily chosen 2D anchors with smoothing factor (left) $\sigma^2 = 1$ and (right) $\sigma^2 = 9$. (Bottom) Membership probabilities for 1D anchors for (left) equation 4 with $\sigma^2 = 0.8$ and (right) equation 2 with $p_1 = p_2 = p_3$, $\sigma_1^2 = 0.04$, and $\sigma_2^2 = \sigma_3^2 = 0.8$. The anchors are landmarked.

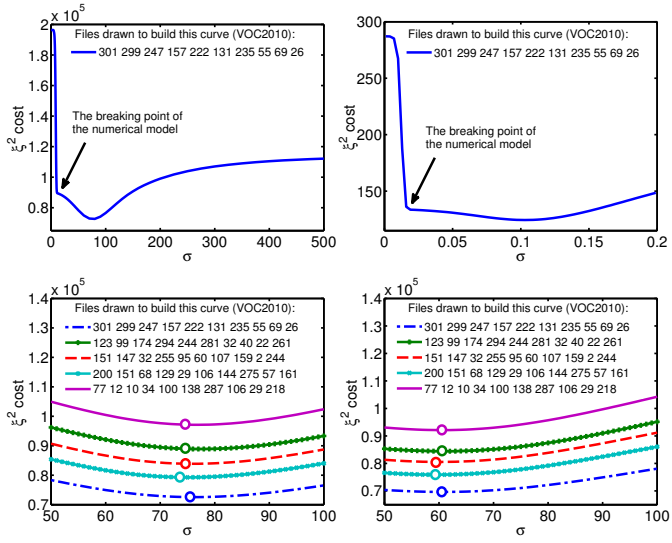

Fig. 2. (Top) Cost function across large range of σ values for grey-SIFT (left) normalised to 255 (RDSS) and (right) with unit length (k-means). (Bottom) Uncertainty of σ on Pascal 2010 Action Classification (left) for RDSS and (right) k-means. Refer text for details.

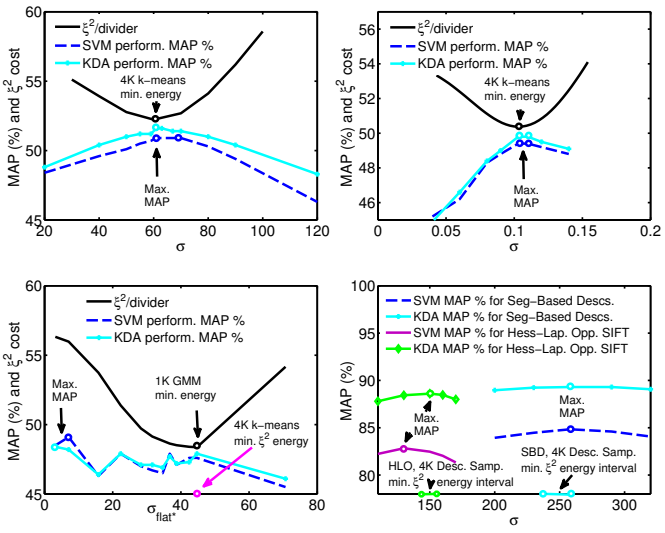

Fig. 3. (Top) MAP maxima and ξ_n^2 minima (VOC2010, k-means, two variants of SIFT, Soft Assignment eq. 4). (Bottom left) MAP maxima and ξ_n^2 for GMM given by equation 2. (Bottom right) MAP maxima and ξ_n^2 minima intervals on Flower17 (RDSS vocabulary, Opp. SIFT, Seg-Based Desc.).

4. EVALUATIONS AND CLASSIFICATION RESULTS

This section provides an experimental insight regarding the quality of the achieved descriptor approximations and classification performance. Tests were performed on Pascal 2010 [8] Action Classification set (301 training, 307 validation, and 613 testing bounding boxes) and Flower 17 [9] set (3 splits of data, each consisting of 680 training, 340 validation, and 340 testing images). For Pascal 2010, we report our results mainly on validation set as testing set is not publicly available. However, we also list test results of our run of the outlined approach submitted for Pascal 2010 competition [8]. Three variants of descriptors were used to scrutinise the behaviour of our cost function. Grey-scale SIFT [10] were extracted on Pascal 2010 with dense feature sampling on a regular grid. Intervals of 8, 14, 20, and 26 pixels, and patch radii of 16, 24, 32, and 40 pixels were applied. This produced 1200 descriptor vectors per image on average. For Flower 17, Opponent SIFT [11] (at Harris Laplace locations) and Segmentation-Based Descriptors [12] were extracted. Both descriptor variants resulted in 2300 vectors per image on average. KDA [13] and SVM classifiers were applied interchangeably to χ^2 [13] and linear kernels formed from SA histograms optimised according to our scheme as in section 3. Spatial Pyramid Match (SPM) [14] with 3 levels of depth was employed. The dictionaries (typically $K = 4000$ anchors) were produced on training sets by either randomly sampling the descriptors (Random Descriptor Set Sampling aka RDSS), by k-means, or by estimating GMM parameters as in equation 3.

First, we provide with an empirical glimpse at the convexity of ξ^2 cost in equation 6 with respect to σ. 10 training images were drawn at random from Pascal 2010 set as this suffices to estimate σ well for the whole set. Both RDSS and k-means were experimented with. Next, the reconstruction error was evaluated as a function of the smoothing factor within a very large range. In figure 2, one can see the cost curves for grey-scale SIFT [10] with vectors normalised to length 255 (top left) and 1 (top right). RDSS and k-means vocabularies were applied respectively. The produced quantisation error curves have several interesting properties: i) to the left of the breaking point (low σ) the numerical accuracy is insufficient to compute the ratio of Gaussians in equation 4, ii) this point can be considered as an approximation to Hard Assignment due to the lowest tangible value of σ iii) there exists a unique minimum, and iv) as $\sigma \rightarrow \infty$, the reconstruction error tends to a value corresponding to the total blurring: all descriptors are equally assigned to all K anchors.

Figure 2 illustrates how much the estimated σ varies with a subset of the drawn descriptors for RDSS (left) and k-means (right) vocabularies. Five-fold drawing process was employed, each time 10 unique descriptor files (one per image) were picked at random. Despite different absolute values of the energy, the minima are located roughly at the same position (negligible uncertainty). Note k-means achieved lower energies. Also, optimal σ for k-means and RDSS differs.

Figure 3 (top) presents MAP performance and σ estimation for k-means vocabulary on Pascal 2010 Action Classification. KDA and SVM were applied to χ^2 kernels. Both MAP% and the energy ξ_n^2 were brought to the same scale with 'divider' to reveal strong corelation between extrema of both measures. Optima are marked as bulky dots on curves.

One can see the best classification performance indeed was achieved for σ estimated according to equation 6. Plot 3 (top left) concerns grey-scale SIFT [10] normalised to 255 with Spatial Coordinate Coding [15].

Top right plot is based on grey-scale unit normalised SIFT [10] and SPM (3 levels of depth). We also evaluated RDSS dictionary concluding the computed σ was optimal. Alas, this vocabulary gave lover results by about 0.5% MAP compared to k-means. Soft Assignment (SA) was further compared to Sparse Coding (SC) [3]. The same k-means dictionary was used and 4 levels of SPM to maximise performance of SC. Though, SC yielded only 48.7% whilst SA reached 49.4% MAP.

Figure 3 (bottom left) presents MAP performance and ξ_n^2 achieved by SA with full parameter GMM according to equation 2. The flattening σ^*_{flat} forces all $\sigma_k \leq \sigma^*_{flat}$ to σ^*_{flat}. It was varied to show impact of non-uniform versus uniform σ_k. If majority of σ_k become equalised, ξ_n^2 drops resulting in a better model. Gradually, local MAP maxima align with the minimum of ξ_n^2. Varying components of GMM between 8-100K did not decrease substantially energies nor improve MAP. Due to the issues discussed in section 3 and illustrated on figure 1 (bottom left), optimising ξ_n^2 for full GMM does not offer the correct σ. Only the reduced GMM model in equation 2 follows closely the descriptor reconstruction approach.

Lastly, figure 3 (bottom right) presents MAP and ξ_n^2 minima intervals on Fower17 set with Harris Laplace Opponent SIFT [11] and Segmentation-Based Descriptor [12]. We observed KDA always worked better with χ^2 and SVM with linear kernels, thus we used such set-up. Our optimisation scheme from equation 6 and SA from equation 4 were applied. For Opponent SIFT, the optimum σ varied between 145-155 for different subsets of 10 randomly drawn images. The diversions from the maximum MAP (up to 0.2%) were noted for σ estimations on less than 10 randomly picked images. Similarly, σ estimated on Segmentation Based Descriptors varied between 240-258 with 0.13% uncertanty in MAP.

Our results on Pascal 2010 [8] Action Classification amount to **62.15**% MAP (average over APs of all 9 concepts on testing set) outperforming other reported systems [8]. They were achieved by averaging multiple kernels of different descriptor variants as in [12, 13]. Flower17 [9] resulted in **89.3**% MAP (**85.4**% accuracy) using Segmentation-Based Descriptor. Multikernel learning [16] yields 86.7% accuracy.

5. CONCLUSIONS

We have presented a novel method for optimising the smoothing factor σ for Visual Word Uncertainty (Soft Assignment). It is extensively demonstrated that the reconstruction error ξ_n^2 has strong impact on the classification performance. We have discussed relation of Soft Assignment to Linear Coordinate Coding methods. Further, we demonstrated why standard GMM cannot perform well in the descriptor reconstruction scenario. Our endeavours led us to state-of-the-art results on both Pascal 2010 Action Classification and Flower17 sets.

Acknowledgements. This work was sponsored by the BBC Future Media and Technology and EPSRC EP/F003420/1 research grants. We would like to also thank Mark Barnard and Fei Yan for several insightful discussions.

6. REFERENCES

[1] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek, "Visual word ambiguity," *PAMI*, 2010.

[2] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," *NIPS*, 2009.

[3] J. Yang, K. Yu, Y. Gong, and T. S. Huang, "Linear spatial pyramid matching using sparse coding for image classification.," *CVPR*, pp. 1794–1801, 2009.

[4] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," *CVPR*, 2010.

[5] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," *CVPR*, vol. 0, pp. 1–8, 2007.

[6] X. Zhou, K. Yu, T. Zhang, and T. S. Huang, "Image classification using super-vector coding of local image descriptors," *ECCV*, pp. 141–154, 2010.

[7] J. Bilmes, "A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models," Tech. Rep., 1998.

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

[9] M. E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," *ICCV*, pp. 722–729, 2008.

[10] D. G. Lowe, "Object recognition from local scale-invariant features," *CVPR*, vol. 2, pp. 1150–1157, 1999.

[11] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "A comparison of color features for visual concept classification," *CIVR*, pp. 141–149, July 2008.

[12] P. Koniusz and K. Mikolajczyk, "On a quest for image descriptors based on unsupervised segmentation maps," *ICPR*, pp. 762–765, 2010.

[13] M.A. Tahir, J. Kittler, K. Mikolajczyk, F. Yan, K. van de Sande, and T. Gevers, "Visual category recognition using spectral regression and kernel discriminant analysis," *ICCV Workshop*, 2009.

[14] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," *CVPR*, vol. 2, pp. 2169–2178, 2006.

[15] P. Koniusz and K. Mikolajczyk, "Spatial coordinate coding to reduce histogram representations, dominant angle and colour pyramid match," *ICIP*, 2011.

[16] F. Yan, K. Mikolajczyk, M. Barnard, H. Cai, and J. Kittler, "Lp norm multiple kernel fisher discriminant analysis for object and image categorisation," *CVPR*, 2010.