Object Recognition Using Hybrid Boosting Method

Osama Ashfaq
Bahria University

Abstract

Li proposed a novel generative/discriminative way to combine features with different types and use them to learn labels in the images. However, the mixture of Gaussian used in Li's paper sufferes greatly from the curse of dimensionality. Here I propose an alternative approach to generate local region descriptor. I treat GMM with diagonal covariance matrix and PCA as separate features, and combine them as the local descriptor. In this way, we could reduce the computational time for mixture model greatly while score greater 90% accuracies for caltech-4 image sets.

1 Introduction

Data science has been vastly changing the world. While human could recognize objects in images with litter effort, object recognition by computer is one of the most challenging tasks in machine learning and computer vision. It has been widely used in technology industries, for example, automatic face extraction and recognition [1, 2]. Like other pattern recognition problems, object recognition aims to classify images (patterns) based on its appearances or other feature extracted from the images. It consists of a feature extraction phase that computes the low dimensional feature descriptors from the high dimensional images, and a classification phase that classify the images relying on the feature descriptors.

Images themselves lie in a very high dimensional space, for example, a 512-by-512 greytone image lies in a $262,144$ dimensional space. Therefore, instead of the whole image, statistical information based on image regions are used for recognizing object classes [3, 4, 5]. However, various segmentation methods produce a variable number of regions. Moreover, within each region, local descriptors based on different feature types produce vectors with various lengths. For instance, a color feature descriptor computes the 3-d mean RGB values in the region while, a SIFT descriptor computes the 128-d histogram of orientations in same region. It is then difficult to combine multiple feature types and arbitrary number of local feature vectors with variable lengths into a single image descriptor. Y. Li [6, 7] proposed an elegant generative/discriminative learning procedure first produce a fixed length image descriptor that summarizes those feature information. Standard supervised learning classifiers are then used to classify the labels of images represented by this fixed length description.

For example, support vector machine (SVM) [8, 9, 10] minimizes the hinge loss function between the training data and the model prediction. Latent models for sequential data [11, 12, 13, 14] maximizes the conditional likelihood of observed data. Logistic regression minimized the negative log conditional likelihood of training data given the model.

In this project, I developed a classification methodology that based on Li's approach. My approach differs from Yi's in the generative phase. Li used an unsupervised clustering method to normalized the description length of local feature vector. Clustering with a Gaussian mixture model (GMM) suffers seriously from the curse of dimensionality [15, 16]. It needs to estimate $O(d^2)$ parameters for d-dimensional data. To deal with this problem, I first restrict the covariance matrices for the GMM be diagonal. I also use PCA to project the high dimension feature vectors into a low dimensional sub-space. A combination of Gaussian components and PCA projections are then used for

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2220v1 | CC BY 4.0 Open Access | rec: 3 Jul 2016, publ: 3 Jul 2016

the generative phase. At the end, the extracted features are trained with DABoost [17], a popular Boosting algorithm known for its resistance to over-fitting.

2 Algorithm

The learning algorithm used in this project are the follows:

a As shown in Figure 1 for each image I_i, a region detector α such as MSER [18] or scale salient detector [19]. Each region is represented by the SIFT [20] descriptor $X^{\alpha}_{i,r}, \forall r \in \{1, \ldots, n^{\alpha}_i\}$. During this approach, both the set $\{X^{\alpha}_{i,r}\}$ and n^{α}_i are stored. Further learning procedures are only dependent on these two parameters, not independent on I_i.

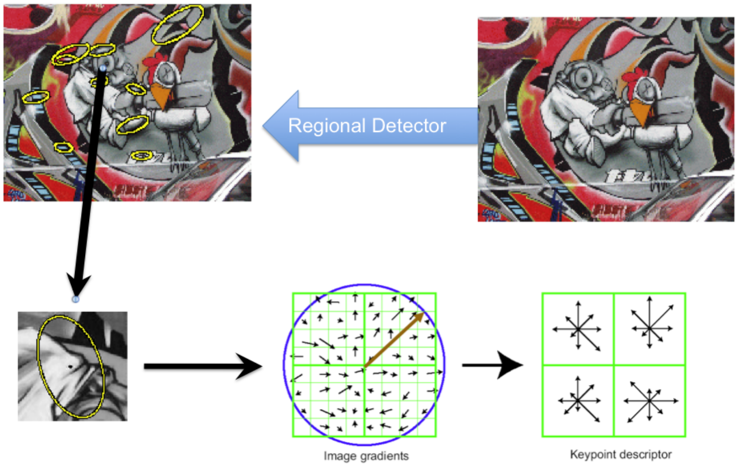


Figure 1: Extract feature vectors from an input image

b Here we only consider one feature type. To reduce the dimensionality of the SIFT descriptor X, the EM algorithm is used to generate a M-component GMM. That is, the EM algorithm approximates the condition probability distribution of X given the object o has the following form,

$$P(X|o) \sim \sum_{m=1}^{M} \omega_m N(X, \mu_m, \Sigma_m) \tag{1}$$

Once the GMM is learned, with ω_m, μ_m, and Σ_m being estimated, the distance between the region descriptor $X_{i,r}$ and cluster centroid μ_m can be calculated. I then use the m-d distance vector to replace the original feature descriptor $X_{i,r}$:

$$\begin{aligned} s(X^{\alpha}_{i,r}, m) &= \log P(X^{\alpha}_{i,r}, m|o) \\ &= \log(\omega_m N(X_{i,r}, \mu_m, \Sigma_m) \end{aligned} \tag{2}$$

For a 128-d SIFT descriptor, EM algorithm needs to estimate $M \times (\frac{128*129}{2} + 128 + 1) = 8385M$ parameters. It is then nearly impossible for the EM algorithm to actually converge. Instead,I restrict the covariance matrix Σ be diagonal, then I only need to estimate $(128 + 128 + 1)M$ parameters. This simplifications speed up the convergence greatly. However, it assumes independent relations between each components of $X_{i,r}$. The diagonal GMM model may be a good approximation to true conditional distribution $P(X_{i,r}|r)$ It is a trade-off between the quality of approximation and the speed.

c Since the information contains in $s(X^{\alpha}_{i,r}, m)$ may not be complete, I seeks for other dimensional reduction techniques to find a low dimensional representation of X.

Figure 2 illustrates the eigen-spectrum of the covariance of $X_{i^o,r}$, where the image I_{i^o} contains object o. As indicated by the points within the red circle, there are about $8-$

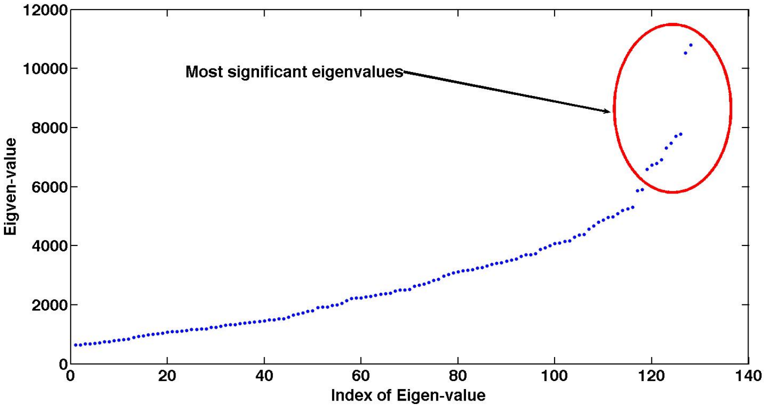


Figure 2: Eigen-spectrum of $cov(X_{i^o,r})$

10 most significant eigen-vectors that have eigenvalues noticeably greater than the other. Alone these eigen-dimensions, $X_{i^o,r}$ varies the most. Therefore, we can project the SIFT feature $X_{i^o,r}$ to the eigen space spanned by $\nu_{k=1}^K$ where ν_k are the most K significant eigen-vectors.

$$q(X_{i,r},k) = (X_{i,r} - \bar{X})\nu_k \quad \forall k \in 1,\dots,K \tag{3}$$

The combined region descriptor then has the form $f(X_{i,r}) = [s(X_{i,r}^\alpha, m), q(X_{i,r},k)|m = 1\dots M, k = 1\dots K]$ with length $M + K$.

d For feature type α, $f(X_{i,r}^\alpha, j)$ is the local region descriptor, an aggregated image descriptor is computed from $f(X_{i,r}^\alpha, j)$:

$$F^\alpha(I_i, j) = \max_r\{f(X_{i,r}^\alpha, j)|r = 1\dots n_i\} \tag{4}$$

In this way, the number of features may vary from one image to another. The length of image descriptor F is still fixed.

e For a different feature type α', repeat step (b) to (d) to get another image descriptor $F^{\alpha'}$. A simple concatenation of F^α and $F^{\alpha'}$ will give us the new combined image descriptor.

f Assign label 1 to images that contain object o and 0 to those do not contain. Train a supervised learner like linear SVM or multilayer neural net with image descriptor and image label.

3 Experiments

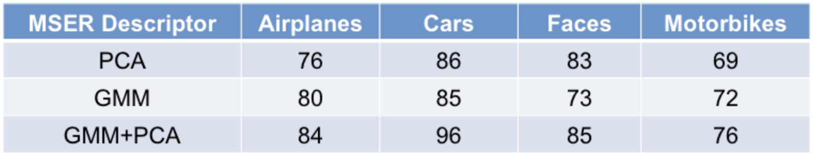
In experiments, Caltech-4 [21] (airplanes, cars, faces, and motorbikes) data set was used. For each category, 200 images are selected. I learn one object class at a time. During learning, the training set consists of 100 positive examples and 100 negative examples from the other categories.

First I use MSER [18] as region detector and two-layer neural net as classifier. The number of Gaussian components is $M = 8$, and the dimension of eigenspace is also $K = 8$. Figure 3 shows the prediction accuracy for the my learning procedure. As it illustrated, GMM alone or PCA projection alone is not sufficient to learn the label well. However, the combination of the two descriptors helps classification a lot. We could achieve 96% classification accuracy for cars, about 85% for airplane and faces. However, MSER detector doesn't handle motorbike quite well, only score 75% accuracy.

As suggested in Li's paper [6], I also implement Kadir's scale salient [19] detector for better classification of motorbikes. The learning results for using Kadir's detector alone are shown figure 4. The number of Gaussian components and the dimension of eigen-space stay the same. As the figure shows, Kadir is 10% better in recognizing motorbikes, but 4% worst in airplanes.

A combination of MSER and Kadir descriptor (with length = 32) gives us the best of both worlds. The learning accuracies are all above 90%, as shown in figure 5.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2220v1 | CC BY 4.0 Open Access | rec: 3 Jul 2016, publ: 3 Jul 2016



MSER Descriptor	Airplanes	Cars	Faces	Motorbikes
PCA	76	86	83	69
GMM	80	85	73	72
GMM+PCA	84	96	85	76

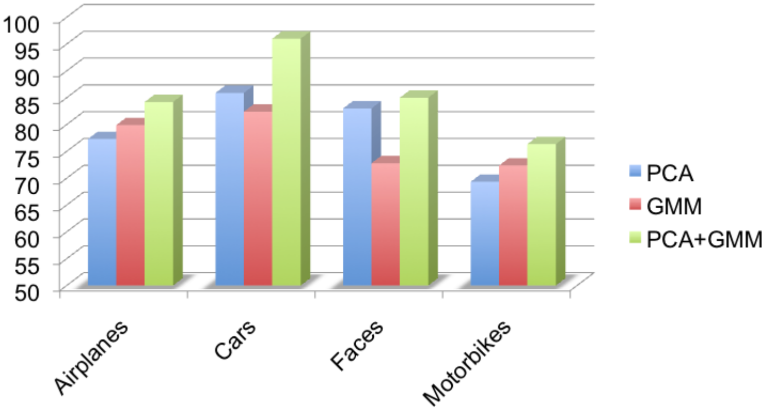
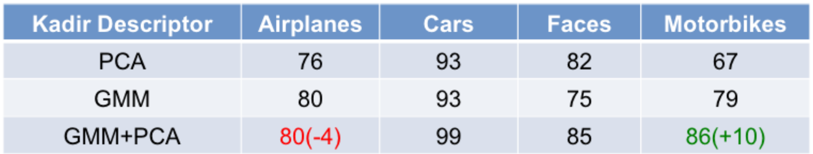


Figure 3: Prediction accuracy for MSER region detector, in percentage.



Kadir Descriptor	Airplanes	Cars	Faces	Motorbikes
PCA	76	93	82	67
GMM	80	93	75	79
GMM+PCA	80(-4)	99	85	86(+10)

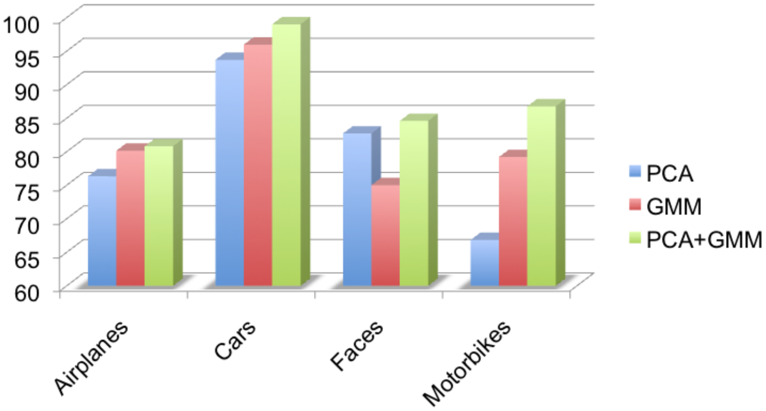


Figure 4: Prediction accuracy for Kadir's region detector, in percentage.

4 Conclusion and future work

In this project, I use both Gaussian mixture model and PCA to reduce the dimensionality of the SIFT descriptor. Following Li's [6, 17] approach, an aggregated image descriptor is calculated based on the local feature vector. And an unsupervised learner then classifies images based on that fixed length image descriptor.

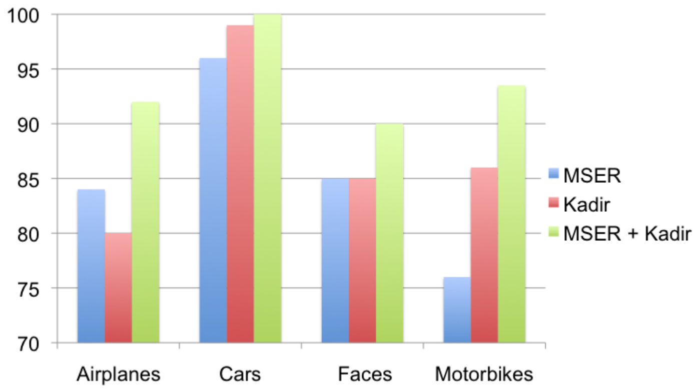


Figure 5: Prediction accuracy for using both MSER and Kadir's region detector, in percentage.

In the experiments, I observed that feature descriptor based Kadir's regions are in general good at motorbikes while that based on MSER regions are better at airplanes. And PCA projections will improve the recognition rate for faces significantly. That may be the reason why so many people use eigen-faces.

I also observe that the performance of learner increase as the size of learning set expands, as predicted by the learning theory. Another interesting finding is that the performance becomes better as we increase the length of descriptor. When using GMM or PCA alone, the performance is just so so. A combination of GMM and PCA improve the recognition a bit. A further combination of features based on Kadir and MSER regions even push the accuracy to the limit. One might argued that the increase in performance may be resulted from increase in description length, not a result from combining multiple features. However, even we could observed an performance increase if we increase the number of Gaussian components to 16 or 32, when we use GMM alone. However, it took much longer time to learn GMM with $M = 16$ or $M = 32$. Put it another way, the sum of the time we use to compute the GMM with $M = 8$ based on MSER, and the time we use compute the same GMM with $M = 8$ but based on Kadir, is much less than the time we spend on learning a single GMM with $M = 16$.

I also use SIFT descriptor in this projecct, a future extension to this project may include develop a simple model for the spatial relationships among the parts, or a model based on color and texture information, and train the system to recognize objects according to both parts and relationships

References

[1] J Wu, R Tse, CL Heike, and LG Shapiro. Learning to compute the symmetry plane for human faces. In *ACM Conference on Bioinformatics, Computational Biology and Health Informatics*, 2011.

[2] Jia Wu, Raymond Tse, and Linda G Shapiro. Automated face extraction and normalization of 3d mesh data. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE*, pages 750–753. IEEE, 2014.

[3] O. Maron and AL.L. Ratan. Multiple-instance learning for natural scene classification. *ICML*, 1998.

[4] Y.Li and L. G. Shapiro. Consistent line clusters for building recognition in cbir. *ICPR*, 2002.

[5] Sanjiv Kumar, Alexander C. Loui, Er C. Loui B, and Martial Hebert. An observation-constrained generative approach for probabilistic classification of image regions. *Image and Vison Computing*, 2003.

[6] Y. Li, L.G. Linda, and J. Bilmes. A generative/discriminative learning algorithm for image classification. *ICCV*, 2005.

[7] Congle Zhang, Raphael Hoffmann, and Daniel S Weld. Ontological smoothing for relation extraction with minimal supervision. In *AAAI*, 2012.

[8] Marti A. Hearst, Susan T Dumais, Edgar Osman, John Platt, and Bernhard Scholkopf. Support vector machines. *Intelligent Systems and their Applications, IEEE*, 13(4):18–28, 1998.

[9] Bernhard Schölkopf and Alex Smola. Support vector machines. *Encyclopedia of Biostatistics*, 1998.

[10] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[11] Lawrence R Rabiner and Biing-Hwang Juang. An introduction to hidden markov models. *ASSP Magazine, IEEE*, 3(1):4–16, 1986.

[12] LR Bahl, Peter F Brown, Peter V De Souza, and Robert L Mercer. Maximum mutual information estimation of hidden markov model parameters for speech recognition. In *proc. icassp*, volume 86, pages 49–52, 1986.

[13] Yanping Huang and Rajesh P Rao. Neurons as monte carlo samplers: Bayesian inference and learning in spiking networks. In *Advances in Neural Information Processing Systems 27*, pages 1943–1951. 2014.

[14] Yanping Huang and Rajesh P.N. Rao. Bayesian inference and online learning in poisson neuronal networks. *Neural Computation*, 28(8), 2016.

[15] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification*. 2002.

[16] Congle Zhang, Tyler Baldwin, Howard Ho, Benny Kimelfeld, and Yunyao Li. Adaptive parser-centric text normalization. In *ACL (1)*, pages 1159–1168, 2013.

[17] Congle Zhang and Daniel S Weld. Harvesting parallel news streams to generate paraphrases of event relations. In *EMNLP*, pages 1776–1786, 2013.

[18] J.Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide baseline stereo from maximally stable extremal regions. *BMVC*, 2002.

[19] T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. *IJCV*, 2004.

[20] D.Lowe. Distinctive image features from scale invariant keypoints, 2004.

[21] Computational vision at caltech. http://www.vision.caltech.edu/archive.html.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2220v1 | CC BY 4.0 Open Access | rec: 3 Jul 2016, publ: 3 Jul 2016