Fisher vector image representation

Jakob Verbeek
January 13, 2012

Course website:
Fisher vector representation

- Alternative to bag-of-words image representation introduced in *Fisher kernels on visual vocabularies for image categorization* 

- FV in comparison to the BoW representation
  - Both FV and BoW are based on a visual vocabulary, with assignment of patches to visual words
  - FV based on Mixture of Gaussian clustering of patches, BoW based on k-means clustering
  - FV Extracts a larger image signature than the BoW representation for a given number of visual words
  - Leads to good classification results using linear classifiers, where BoW representations require non-linear classifiers.
Fisher vector representation: Motivation 1

• Suppose we use a bag-of-words image representation
  – Visual vocabulary trained offline

• Feature vector quantization is computationally expensive in practice

• To extract visual word histogram for a new image
  – Compute distance of each local descriptor to each k-means center
  – Run-time $O(NKD)$: linear in
    • $N$: nr. of feature vectors $\sim 10^4$ per image
    • $K$: nr. of clusters $\sim 10^3$ for recognition
    • $D$: nr. of dimensions $\sim 10^2$ (SIFT)

• So in total in the order of $10^9$ multiplications per image to obtain a histogram of size 1000

• Can this be done more efficiently ?!
  – Yes, extract more than just a visual word histogram!
Fisher vector representation: Motivation 2

• Suppose we want to refine a given visual vocabulary

• Bag-of-word histogram stores # patches assigned to each word
  – Need more words to refine the representation
  – But this directly increases the computational cost
  – And leads to many empty bins, redundancy
Fisher vector representation: Motivation 2

• Instead, the Fisher Vector also records the mean and variance of the points per dimension in each cell
  – More information for same # visual words
  – Does not increase computational time significantly
  – Leads to high-dimensional feature vectors

• Even when the counts are the same the position and variance of the points in the cell can vary
Image representation using Fisher kernels

• General idea of Fischer vector representation
  – Fit probabilistic model to data \( p(X; \Theta) \)
  – Represent data with derivative of data log-likelihood
    “How does the data want that the model changes?”
    \[
    G(X, \Theta) = \frac{\partial \log p(x; \Theta)}{\partial \Theta}
    \]
    
    Jaakkola & Haussler. “Exploiting generative models in discriminative classifiers”,

• We use Mixture of Gaussians to model the local (SIFT) descriptors \( X = \{x_n\}_{n=1}^N \)
  \[
  L(X, \Theta) = \sum_n \log p(x_n)
  
  p(x_n) = \sum_k \pi_k \, N(x_n; m_k, C_k)
  \]
  – Define mixing weights using the soft-max function
  \[
  \pi_k = \frac{\exp \alpha_k}{\sum_k \exp \alpha_k},
  \]
  ensures positiveness and sum to one constraint
Image representation using Fisher kernels

- Mixture of Gaussians to model the local (SIFT) descriptors
  \[ L(\Theta) = \sum_n \log p(x_n) \]
  \[ p(x_n) = \sum_k \pi_k N(x_n; m_k, C_k) \]
  - The parameters of the model are \( \Theta = (\alpha_k, m_k, C_k)_{k=1}^K \)
  - where we use diagonal covariance matrices

- Concatenate derivatives to obtain data representation
  \[ G(X, \Theta) = \left( \frac{\partial L}{\partial \alpha_1}, \ldots, \frac{\partial L}{\partial \alpha_K}, \frac{\partial L}{\partial m_1}, \ldots, \frac{\partial L}{\partial m_K}, \frac{\partial L}{\partial C_1^{-1}}, \ldots, \frac{\partial L}{\partial C_K^{-1}} \right)^T \]
Image representation using Fisher kernels

- Data representation

\[ G(X, \Theta) = \left( \frac{\partial L}{\partial \alpha_1}, \ldots, \frac{\partial L}{\partial \alpha_K}, \frac{\partial L}{\partial m_1}, \ldots, \frac{\partial L}{\partial m_K}, \frac{\partial L}{\partial C_1^{-1}}, \ldots, \frac{\partial L}{\partial C_K^{-1}} \right)^T \]

- In total \( K(1+2D) \) dimensional representation, since for each visual word / Gaussian we have

  \[ \frac{\partial L}{\partial \alpha_k} = \sum_n (q_{nk} - \pi_k) \]  

  \[ \frac{\partial L}{\partial m_k} = C_k^{-1} \sum_n q_{nk} (x_n - m_k) \]  

  \[ \frac{\partial L}{\partial C_k^{-1}} = \frac{1}{2} \sum_n q_{nk} (C_k - (x_n - m_k)^2) \]

  With the soft-assignments: \[ q_{nk} = p(k|x_n) = \frac{\pi_k p(x_n|k)}{p(x_n)} \]
Bag-of-words vs. Fisher vector image representation

- Bag-of-words image representation
  - Off-line: fit k-means clustering to local descriptors
  - Represent image with histogram of visual word counts: $K$ dimensions

- Fischer vector image representation
  - Off-line: fit MoG model to local descriptors
  - Represent image with derivative of log-likelihood: $K(2D+1)$ dimensions

- Computational cost similar:
  - Both compare $N$ descriptors to $K$ visual words (centers / Gaussians)

- Memory usage: higher for fisher vectors
  - Fisher vector is a factor $(2D+1)$ larger, e.g. a factor 257 for SIFTs!
    - Ie for 1000 visual words this is roughly $257 \times 1000 \times 4$ bytes $\sim$ 1 Mb
  - However, because we store more information per visual word, we can generally obtain same or better performance with far less visual words
Images from categorization task PASCAL VOC

- Yearly evaluation since 2005 for image classification (also object localization, segmentation, and body-part localization)
Fisher vectors: classification performance

- Results taken from: “Fisher Kernels on Visual Vocabularies for Image Categorization”, F. Perronnin and C. Dance, in CVPR '07

- BoW and Fisher vector yield similar performance
  - Fisher vector uses 32x fewer Gaussians
  - BoW representation 2,000 long, FV length is 64(1+2 x 128) = 16,448
Additional reading material

• Fisher vector image representation
  – “Fisher Kernels on Visual Vocabularies for Image Categorization”
    F. Perronnin and C. Dance, in CVPR '07

• Pattern Recognition and Machine Learning
  Chris Bishop, 2006, Springer
  - Section 6.2
Exam

• Friday January 27th
  – From 9 am to 12 am
  – Room H105 Ensimag building @ campus

• Prepare from
  – Lecture slides
  – Presented papers
  – Bishop's book

• During the exam you can bring
  – the lecture slides
  – the presented papers