Bag-of-words model in computer vision

In **computer vision**, the **bag-of-words model** (BoW model) sometimes called **bag-of-visual-words model** [1][2] can be applied to image classification or retrieval, by treating image features as words. In document classification, a bag of words is a sparse vector of occurrence counts of words; that is, a sparse histogram over the vocabulary. In computer vision, a *bag of visual words* is a vector of occurrence counts of a vocabulary of local image features.

Image representation based on the BoW model

To represent an image using the BoW model, an image can be treated as a document. Similarly, "words" in images need to be defined too. To achieve this, it usually includes following three steps: feature detection, feature description, and codebook generation. [1][2][3] A definition of the BoW model can be the "histogram representation based on independent features".[4] Content based image indexing and retrieval (CBIR) appears to be the early adopter of this image representation technique.[5]

Feature representation

After feature detection, each image is abstracted by several local patches. Feature representation methods deal with how to represent the patches as numerical vectors. These vectors are called feature descriptors. A good descriptor should have the ability to handle intensity, rotation, scale and affine variations to some extent. One of the most famous descriptors is Scale-invariant feature transform (SIFT).[6] SIFT converts each patch to 128-dimensional vector. After this step, each image is a collection of vectors of the same dimension (128 for SIFT), where the order of different vectors is of no importance.

Codebook generation

The final step for the BoW model is to convert vector-represented patches to "codewords" (analogous to words in text documents), which also produces a "codebook" (analogy to a word dictionary). A codeword can be considered as a representative of several similar patches. One simple method is performing k-means clustering over all the vectors.[7] Codewords are then defined as the centers of the learned clusters. The number of the clusters is the codebook size (analogous to the size of the word dictionary).

Thus, each patch in an image is mapped to a certain codeword through the clustering process and the image can be represented by the histogram of the codewords.

Learning and recognition based on the BoW model