In this paper, we propose an approach to improve image retrieval performance. As shown in following figure, Query image \( q \) has three Visual Words. None of them has a match in database image \( p \). Previous methods based on inner product or cosine measure will give a zero score for this match. With the proposed method, we consider the neighboring sets and thus it may result in a non-zero matching score.

Math Representations and Methodologies

- Image is represented by SIFT features of 128 dimensions. All SIFT features \( f_{ij}, i = 1...N, j = 1...M_i \), are clustered into different clusters (calculated visual word) \( v_{w_i}, k = 1...K \).
- Each image is represented by a ‘Bag of Visual Words’ (BVW). Retrieval is done by comparing BVW of query image \( q \) with BVW of database image \( p \).
- Existing methods: if Visual Word \( v_{w} \in \text{BVW}_q \) does not have a match in BVW, it will not receive a zero score in the retrieval process. In reality, Visual Word \( v_{w} \) may have a very close neighbor in BVW. So, previous methods fail to take this case into account.
- In proposed method, we consider all neighbors \( \text{NS}(v_{w}) \) of Visual Words \( v_{w} \) in query image. Even if there is no match for \( v_{w} \) in a database image BVW, it is still possible to have a non-zero match.

Constructing Neighboring Sets of Visual Words

- We construct neighboring sets based on distances between Visual Words. We have several considerations.
  - If two visual words are close enough, we use the Euclidean distance \( \text{EUC}(.) \) as their distance measure.
  - If two visual words are further than a threshold \( T_i \), we take the shortest path distance \( \text{SP}(.) \) as their distance measure.
  - If two visual words are further than a threshold \( T_h \), we set their distance to \( \infty \).
- Distance between Visual Words
  
  \[
  \text{Dis}(v_{w_i}, v_{w_j}) = \begin{cases} 
  \text{EUC}(v_{w_i}, v_{w_j}) & \text{if } \text{EUC}(v_{w_i}, v_{w_j}) < T_i \\
  \text{SP}(v_{w_i}, v_{w_j}) & \text{if } T_i < \text{EUC}(v_{w_i}, v_{w_j}) < T_h \\
  \infty & \text{if } \text{EUC}(v_{w_i}, v_{w_j}) > T_h
  \end{cases}
  \]
- Constructing neighboring set
  
  \( \text{NS}(v_{w_i}) = \{v_{w_j} | \text{Dis}(v_{w_i}, v_{w_j}) < \infty\} \)

Normalized Distances

- We introduce a normalized distance between two visual words in the same neighboring sets:
  
  \[
  \text{D}(v_{w_i}, v_{w_j}) = \frac{\text{Dis}(v_{w_i}, v_{w_j})}{\sum_{0<j<k} \text{Dis}(v_{w_i}, v_{w_j})}
  \]

Proposed Image Retrieval Algorithm

- Query image \( q \):
  - feature points \( f_{iq}, i = 1...n_q \)
  - visual words \( v_{w_i}, i = 1...n_q \)
  - Neighboring Set \( \text{NS}(v_{w_i}) \)
- Image \( p \) in database
  - BVW = \{\( v_{w_i} | i = 1...n_p \)\}
- We propose:
  - Each query feature \( f_{iq} \) receive a score on image \( p \) in the database
  
  \[
  \text{Score}(f_{iq}, p) = \begin{cases} 
  1 & \text{if } v_{w_i} \in \text{BVW}_p \\
  0 & \text{if } \text{NS}(v_{w_i}) \cap \text{BVW}_p = \emptyset \\
  1 - C & \text{otherwise}
  \end{cases}
  \]
  
  where \( C = \sum_{v_{w_j} \in \text{NS}(v_{w_i}) \cap \text{BVW}_p} \text{D}(v_{w_i}, v_{w_j}) \).
  - Proposed similarity measure of query image \( q \) on image \( p \)
  
  \[
  \text{Simi}(q, p) = \frac{\sum_{0 \leq c < n} \text{Score}(f_{iq}, p)}{n_q}
  \]

Experimental Setup

- We used the first 50 categories of Caltech 101 database. The number of images in each category ranges from 33 (binocular) to 800 (airplanes).
- The performance is evaluated using the precision metric. For an input image \( l_i, i = 1...n_c \) in category \( c \), count the number of top \( m \) returns in correct category \( c \) and represent this number as \( n_{c, \text{correct}} \). Then the precision for image \( l_i \) is
  
  \[
  P(c, i, m) = \frac{n_{c, \text{correct}}}{n_c}
  \]
  - The average precision for a category \( c \) at parameter \( m \) is
  
  \[
  \text{AP}(c, m) = \frac{\sum_{0 \leq c < n} P(c, i, m)}{n_c}
  \]

Results: Retrieval Performance

The leftmost one is the query image. The eight images on the right are the returns. Only the 5th image is not in the “accordion” category.

Normalized Distances

- We introduce a normalized distance between two visual words in the same neighboring sets:
  
  \[
  \text{D}(v_{w_i}, v_{w_j}) = \frac{\text{Dis}(v_{w_i}, v_{w_j})}{\sum_{0<j<k} \text{Dis}(v_{w_i}, v_{w_j})}
  \]

Conclusions

- We use the neighboring set of visual words instead of individual visual words in the retrieval process.
- We use the shortest path distance instead of direct Euclidean distance for measuring distance between Visual Words, in case the Visual Word subspace is a non-linear manifold.
- We improved performance over state of the art unsupervised image retrieval.