

$$\begin{aligned}
D(I_q, I_m) &= \sum_{i=1}^N |q_i - m_i|^p \quad (5) \\
&= 2 + \sum_{i|q_i \neq 0, m_i \neq 0} (|q_i - m_i|^p - q_i^p - m_i^p) \quad (6)
\end{aligned}$$

However, the dissimilarity measure by the  $L_p$ -distance is not optimal. As revealed in [182], there exists the neighborhood reversibility issue, which means that an image is usually not the  $k$ -nearest neighbor of its  $k$ -nearest neighbor images. Such issue causes that problem that some images are frequently returned while others are rarely returned when submitting query images. To address this problem, Jegou *et al.* proposed a novel contextual dissimilarity measure to refine the Euclidean distance based distance [182]. It modifies the neighborhood structure in the BoW space by iteratively estimating distance update terms in the spirit of Sinkhorn's scaling algorithm. Alternatively, in [183], a probabilistic framework is proposed to model the feature to feature similarity measure and a query adaptive similarity is derived. Different from the above approaches, in [184], the similarity metric is implicitly learnt with diffusion processes by exploring the affinity graphs to capture the intrinsic manifold of database images.

In [138], Jegou *et al.* investigated the phenomenon of co-missing and co-occurrence in the regular BoW vector representation. The co-missing phenomenon denotes a negative evidence, *i.e.*, a visual word is jointly missing from two BoW vectors. To include the under-estimated evidence for similarity measurement refinement, vectors of images are centered by mean subtraction [138]. On the other hand, the co-occurrence of visual words across BoW vectors will cause over-counting of some visual patterns. To limit this impact, a whitening operation is introduced to the BoW vector to generate a new representation [138]. Such preprocessing also applies to the VLAD vector [116]. Considerable accuracy gain has been demonstrated with the above operations.

inconsistency of the matching between two bundled features. In [185][186], Zheng *et al.* propose a novel  $L_p$ -norm IDF to extend the classic IDF weighting scheme.

The context clues in the descriptor space and the spatial domain are important to contribute the similarity score when comparing images. In [123], a contextual weighting scheme is introduced to enhance the original IDF-based voting so as to improve the classic vocabulary tree approach. Two kinds of weighting scheme, *i.e.*, descriptor contextual weighting (DCW) and spatial contextual weighting, are formulated to multiply the basic IDF weight as a new weighting scheme for image scoring. In [187], Shen *et al.* proposed a spatially-constrained similarity measure based on a certain transformation to formulate voting score. The transformation space is discretized and a voting map is generated based on the relative locations of matched features to determine the optimal transformation.

In [179], each indexed feature is embedded with a binary signature and the image distance is defined as a summation of the hamming distance between matched features, of which the distance for the unobserved match is set as statistical expectation of the distance. Similar scoring scheme for the unobserved match is also adopted by Liu *et al.* [157]. In [63], to tolerate the correspondences of multiple visual objects with different transformations, local similarity of deformations is derived from the peak value in the histogram of pairwise geometric consistency [188]. This similarity score is used as a weighting term to the general voting scores from local correspondences.

In image retrieval with visual word representation, similar to text-based information retrieval [189], there is a phenomenon of visual word burstiness, *i.e.*, some visual element appears much more frequently in an image than the statistically expectation, which undermines the visual similarity measure. To address this problem, Jegou *et al.* proposed three strategies to penalize the voting scores from the bursting visual words by removing multiple local matches and weaken the influence of intra- and inter-images bursts [190] [191].