

## Chapter 5

# Lower Bounds on the EMD

In the content-based retrieval systems described in [65] and [69], the distance between two images is taken as the EMD between the two corresponding signatures. The query time is dominated by the time to perform the EMD computations. In a nearest neighbor query, the system returns the  $R$  database images which are closest to the given query. During query processing, an exact EMD computation need not be performed if there is a lower bound on the EMD which is greater than the  $R$ th smallest distance seen so far. The goal is to perform queries in time that grows in a sublinear fashion with the number of database images. The motivation is system scalability to very large databases.

It is known ([68]) that the distance between the centroids of two equal-weight distributions is a lower bound on the EMD between the distributions if the ground distance is induced by a vector norm. There are, however, common situations in which distributions will have unequal total weights. For example, consider once again the color-based retrieval work described in [65]. Assuming all the pixels in an image are classified, the weight of every database signature is one. EMD comparisons between unequal-weight distributions arise whenever the system is presented with a *partial* query such as: "give me all images with at least 20% sky blue and 30% green". The query signature consists of two points in CIE-Lab space with weights equal to 0.20 and 0.30, and therefore has total weight equal to 0.50. Since one cannot assume that all database images and queries will contain the same amount of information, lower bounds on the EMD between unequal-weight distributions may be quite useful in retrieval systems.

This chapter is organized as follows. In section 5.1, we extend the centroid-distance lower bound to the case of unequal-weight distributions. In section 5.2, we present lower bounds which use projections of distribution points onto random lines through the origin and along the directions of the axes. In section 5.3, we show some experiments that use our

lower bounds in the previously mentioned color-based image retrieval system.

A preliminary version of most of the material in this chapter is contained in the technical report [12].

## 5.1 Centroid-based Lower Bounds

The centroid  $\bar{\mathbf{x}}$  of the distribution  $\mathbf{x} = (X, w) \in \mathbf{D}^{K,m}$  is defined as

$$\bar{\mathbf{x}} = \frac{\sum_{i=1}^m w_i x_i}{w_\Sigma}.$$

In section 5.1.1 we shall prove that the distance between the centroids of equal-weight distributions is a lower bound on the EMD between the distributions if the ground distance is induced by a vector norm or if  $d = L_2^2$ . There is also, however, a centroid-based lower bound if the distributions are not equal weight. If  $\mathbf{x} = (X, w)$  is heavier than  $\mathbf{y} = (Y, u)$ , then all of the weight in  $\mathbf{y}$  is matched to part of the weight in  $\mathbf{x}$ . The weight in  $\mathbf{x}$  which is matched to  $\mathbf{y}$  by an optimal flow is a sub-distribution  $\mathbf{x}'$  of  $\mathbf{x}$ . Formally, a *sub-distribution*  $\mathbf{x}' = (X', w')$  of  $\mathbf{x} = (X, w) \in \mathbf{D}^{K,m}$ , denoted  $\mathbf{x}' \subset \mathbf{x}$ , is a distribution with  $X' = X$  and  $0 \leq w' \leq w$ :

$$\mathbf{x}' = \{ (x_1, w'_1), \dots, (x_m, w'_m) \} = (X, w') \in \mathbf{D}^{K,m}, \quad 0 \leq w'_j \leq w_j \text{ for } j = 1, \dots, m.$$

In words, the points of a sub-distribution  $\mathbf{x}'$  are the same as the points of  $\mathbf{x}$  and the weights of  $\mathbf{x}'$  are bounded by the weights of  $\mathbf{x}$ . One can visualize a sub-distribution  $\mathbf{x}' \subset \mathbf{x}$  as the result of removing some of the dirt in the piles of dirt in  $\mathbf{x}$ . The minimum distance between the centroid of  $\mathbf{y}$  and the locus of the centroid of sub-distributions of  $\mathbf{x}$  of total weight  $u_\Sigma$  is a lower bound on  $\text{EMD}(\mathbf{x}, \mathbf{y})$ . Details are given in section 5.1.2.

### 5.1.1 Equal-Weight Distributions

Let us first consider the case when the ground distance between points is induced by a vector norm. This is true, for example, if the ground distance is one of the  $L_p$  distances ( $p \geq 1$ ).

**Theorem 6** *Suppose  $\mathbf{x} = (X, w) \in \mathbf{D}^{K,m}$  and  $\mathbf{y} = (Y, u) \in \mathbf{D}^{K,n}$  are distributions of equal total weight  $w_\Sigma = u_\Sigma$ . Then*

$$\text{EMD}^{\|\cdot\|}(\mathbf{x}, \mathbf{y}) \geq \|\bar{\mathbf{x}} \leftrightarrow \bar{\mathbf{y}}\|$$

*if the ground distance  $d(x, y) = \|x \leftrightarrow y\|$  and  $\|\cdot\|$  is a norm.*

**Proof.** The equal-weight requirement implies that for any feasible flow  $F = (f_{ij})$ ,

$$\sum_{i=1}^m f_{ij} = u_j \quad \text{and} \quad (5.1)$$

$$\sum_{j=1}^n f_{ij} = w_i. \quad (5.2)$$

Then

$$\begin{aligned} \left\| \sum_{i=1}^m w_i x_i \Leftrightarrow \sum_{j=1}^n u_j y_j \right\| &= \left\| \sum_{i=1}^m \sum_{j=1}^n f_{ij} x_i \Leftrightarrow \sum_{i=1}^m \sum_{j=1}^n f_{ij} y_j \right\| \quad ((5.1), (5.2)) \\ &= \left\| \sum_{i=1}^m \sum_{j=1}^n f_{ij} (x_i \Leftrightarrow y_j) \right\| \\ &\leq \sum_{i=1}^m \sum_{j=1}^n \|f_{ij} (x_i \Leftrightarrow y_j)\| \quad (\Delta\text{-inequality}) \\ &= \sum_{i=1}^m \sum_{j=1}^n f_{ij} \|x_i \Leftrightarrow y_j\| \quad (f_{ij} \geq 0) \\ \left\| \sum_{i=1}^m w_i x_i \Leftrightarrow \sum_{j=1}^n u_j y_j \right\| &\leq \sum_{i=1}^m \sum_{j=1}^n f_{ij} \|x_i \Leftrightarrow y_j\|. \end{aligned}$$

Dividing both sides of the last inequality by  $w_\Sigma = u_\Sigma$  yields

$$\|\bar{\mathbf{x}} \Leftrightarrow \bar{\mathbf{y}}\| \leq \frac{\sum_{i=1}^m \sum_{j=1}^n f_{ij} \|x_i \Leftrightarrow y_j\|}{w_\Sigma} \quad (5.3)$$

for any feasible flow  $F$ . Replacing  $F$  by a work minimizing flow gives the desired result. ■

The centroid lower bound for the equal-weight case also holds when the ground distance is  $L_2^2$ , despite the fact that the square of the Euclidean norm is *not* itself a norm.

**Theorem 7** Suppose  $\mathbf{x} = (X, w) \in \mathbf{D}^{K,m}$  and  $\mathbf{y} = (Y, u) \in \mathbf{D}^{K,n}$  are distributions of equal total weight  $w_\Sigma = u_\Sigma$ . Then

$$\text{EMD}^{\|\cdot\|_2^2}(\mathbf{x}, \mathbf{y}) \geq \|\bar{\mathbf{x}} \Leftrightarrow \bar{\mathbf{y}}\|_2^2,$$

where the ground distance  $d(x, y) = \|x \Leftrightarrow y\|_2^2$  and  $\|\cdot\|_2$  is the Euclidean norm.

**Proof.** Applying the Cauchy-Schwarz inequality  $(\sum_k a_k^2)(\sum_k b_k^2) \geq (\sum_k a_k b_k)^2$  with  $a_k = \sqrt{f_{ij}}$  and  $b_k = \sqrt{f_{ij}}\|x_i \leftrightarrow y_j\|_2$  gives

$$\left( \sum_{i=1}^m \sum_{j=1}^n f_{ij} \right) \left( \sum_{i=1}^m \sum_{j=1}^n f_{ij} \|x_i \leftrightarrow y_j\|_2^2 \right) \geq \left( \sum_{i=1}^m \sum_{j=1}^n f_{ij} \|x_i \leftrightarrow y_j\|_2 \right)^2 \quad (5.4)$$

for every feasible flow  $F$ . The first factor on the left-hand side of (5.4) is equal to the total weight  $u_\Sigma = w_\Sigma$ . From (5.3) in the proof of Theorem 6, the right-hand side of the inequality is greater than or equal to  $\|\bar{\mathbf{x}} \leftrightarrow \bar{\mathbf{y}}\|_2^2 u_\Sigma^2$ . Combining these facts with (5.4) shows that

$$\frac{\sum_{i=1}^m \sum_{j=1}^n f_{ij} \|x_i \leftrightarrow y_j\|_2^2}{u_\Sigma} \geq \|\bar{\mathbf{x}} \leftrightarrow \bar{\mathbf{y}}\|_2^2$$

for every feasible flow  $F$ . Replacing  $F$  by an optimal feasible flow yields the desired result. ■

### 5.1.2 Unequal-Weight Distributions

Let  $\mathbf{x} = (X, w) \in \mathbf{D}^{K,m}$  and  $\mathbf{y} = (Y, u) \in \mathbf{D}^{K,n}$  be distributions with  $w_\Sigma \geq u_\Sigma$ . In any feasible flow  $F = (f_{ij})$  from  $\mathbf{x}$  to  $\mathbf{y}$ , all of the weight  $u_j$  must be matched to weight in  $\mathbf{x}$  so that  $\sum_{i=1}^m f_{ij} = u_j$ , and the total amount of matched weight is  $\sum_{i=1}^m \sum_{j=1}^n f_{ij} = u_\Sigma$ . Let

$$\mathbf{x}^F = \left\{ \left( x_1, \sum_{j=1}^n f_{1j} \right), \left( x_2, \sum_{j=1}^n f_{2j} \right), \dots, \left( x_m, \sum_{j=1}^n f_{mj} \right) \right\} = (X, w^F).$$

Clearly,  $w_\Sigma^F = u_\Sigma$ . By Theorem 6 in the previous section, we know that

$$\text{EMD}(\mathbf{x}^F, \mathbf{y}) \geq \left\| \bar{\mathbf{x}}^F \leftrightarrow \bar{\mathbf{y}} \right\| \quad (5.5)$$

when the ground distance is induced by a vector norm  $\|\cdot\|$ . Note that Theorem 7 implies that the lower bound (5.5) also holds when  $d = L_2^2$  if we replace  $\|\cdot\|$  by  $\|\cdot\|_2^2$ .

From (5.5), it follows that

$$\text{EMD}(\mathbf{x}^F, \mathbf{y}) \geq \min_{F' \in \mathcal{F}(\mathbf{x}, \mathbf{y})} \left\| \bar{\mathbf{x}}^{F'} \leftrightarrow \bar{\mathbf{y}} \right\|, \quad (5.6)$$

where the minimum is taken over all feasible flows  $F'$  from  $\mathbf{x}$  to  $\mathbf{y}$ . Since (5.6) holds for every feasible flow  $F$  from  $\mathbf{x}$  to  $\mathbf{y}$ , we can replace  $F$  by a work minimizing flow  $F^*$  and obtain

$$\text{EMD}(\mathbf{x}, \mathbf{y}) = \text{EMD}(\mathbf{x}^{F^*}, \mathbf{y}) \geq \min_{F' \in \mathcal{F}(\mathbf{x}, \mathbf{y})} \left\| \bar{\mathbf{x}}^{F'} \leftrightarrow \bar{\mathbf{y}} \right\|. \quad (5.7)$$

The minimum on the right-hand side of the inequality (5.7) can be re-stated as the minimum distance of the centroid of  $\mathbf{y}$  to the centroid of any sub-distribution of  $\mathbf{x}$  of total weight  $u_\Sigma$ :

$$\min_{F' \in \mathcal{F}(\mathbf{x}, \mathbf{y})} \left\| \overline{\mathbf{x}^{F'}} \leftrightarrow \overline{\mathbf{y}} \right\| = \min_{\substack{\mathbf{x}' = (X, w') \subset \mathbf{x} \\ w'_\Sigma = u_\Sigma}} \left\| \overline{\mathbf{x}'} \leftrightarrow \overline{\mathbf{y}} \right\|. \quad (5.8)$$

We now argue that (5.8) holds. Clearly,  $\mathbf{x}^{F'}$  is a sub-distribution of  $\mathbf{x}$  with total weight  $u_\Sigma$  for every  $F' \in \mathcal{F}(\mathbf{x}, \mathbf{y})$ . It remains to argue that any sub-distribution  $\mathbf{x}' \subset \mathbf{x}$  with total weight  $u_\Sigma$  is  $\mathbf{x}^{F'}$  for some  $F' \in \mathcal{F}(\mathbf{x}, \mathbf{y})$ . Let  $F'$  be any feasible flow between the two equal-weight distributions  $\mathbf{x}'$  and  $\mathbf{y}$  (the set of such feasible flows is nonempty). The feasible flows  $\mathcal{F}(\mathbf{x}, \mathbf{y})$  between  $\mathbf{x}$  and  $\mathbf{y}$  are exactly those flows which match all  $u_\Sigma$  of  $\mathbf{y}$ -weight to  $u_\Sigma \leq w_\Sigma$  of  $\mathbf{x}$ -weight. Therefore,  $\mathcal{F}(\mathbf{x}', \mathbf{y}) \subset \mathcal{F}(\mathbf{x}, \mathbf{y})$ , and  $F' \in \mathcal{F}(\mathbf{x}', \mathbf{y}) \Rightarrow F' \in \mathcal{F}(\mathbf{x}, \mathbf{y})$ .

Combining (5.7) and (5.8) gives

$$\text{EMD}(\mathbf{x}, \mathbf{y}) \geq \min_{\substack{\mathbf{x}' = (X, w') \subset \mathbf{x} \\ w'_\Sigma = u_\Sigma}} \left\| \overline{\mathbf{x}'} \leftrightarrow \overline{\mathbf{y}} \right\|. \quad (5.9)$$

In section 5.1.2.1 we show how this minimization problem can be formulated as the minimization of a quadratic function (if  $d = L_2$ ) subject to linear constraints. However, solving this quadratic programming problem is likely to take more time than computing the EMD itself. In section 5.1.2.2 we show how to compute a bounding box for the locus of the centroid of any sub-distribution of  $\mathbf{x}$  of total weight  $u_\Sigma$ . The minimum ground distance from the centroid of  $\mathbf{y}$  to the bounding box is a lower bound of the EMD, although it is obviously not as tight as the lower bound in (5.9).

### 5.1.2.1 The Centroid Lower Bound

Given a distribution  $\mathbf{x} = (X, w) \in \mathbf{D}^{K, m}$ , the locus of the centroid of sub-distributions of  $\mathbf{x}$  of weight  $\alpha w_\Sigma$ ,  $0 < \alpha \leq 1$ , is

$$C^\alpha(\mathbf{x}) = \left\{ \frac{\sum_{i=1}^m \tilde{w}_i x_i}{\tilde{w}_\Sigma} : 0 \leq \tilde{w}_i \leq w_i, 0 < \tilde{w}_\Sigma = \alpha w_\Sigma \right\}.$$

If we let  $v_i = \tilde{w}_i / \tilde{w}_\Sigma$  and  $\hat{w}_i = w_i / (\alpha w_\Sigma)$ , then

$$C^\alpha(\mathbf{x}) = \left\{ \sum_{i=1}^m v_i x_i : 0 \leq v \leq \hat{w} = \frac{1}{\alpha} \frac{w}{w_\Sigma}, v_\Sigma = 1 \right\}$$

or, in terms of matrix multiplication,

$$C^\alpha(\mathbf{x}) = \{ Xv : 0 \leq v \leq \hat{w} = \frac{1}{\alpha} \frac{w}{w_\Sigma}, 1^T v = 1 \}. \quad (5.10)$$

The symbol “1” is overloaded in the constraint  $1^T v = 1$ ; on the left-hand side it is a vector of  $m$  ones, while on the right-hand side it is simply the integer one. It is easy to see from (5.10) that

$$C^{\alpha_1}(\mathbf{x}) \supseteq C^{\alpha_2}(\mathbf{x}) \quad \text{if } \alpha_1 \leq \alpha_2.$$

The locus  $C^\alpha(\mathbf{x})$  is a convex polytope. The intersection of the  $2m$  halfspaces  $v \geq 0$  and  $v \leq \hat{w}$  is a box  $P_1$ . The intersection of  $P_1$  with the hyperplane  $1^T v = 1$  is another convex polytope  $P_2$  of one dimension less. Finally, applying the linear map  $X$  to  $P_2$  gives the convex polytope  $C^\alpha(\mathbf{x})$ . In [4], Bern et al. characterize and provide algorithms to compute the locus  $C_{L,H}(S)$  of the centroid of a set  $S$  of points with approximate weights, where weight  $w_i$  lies in a given interval  $[l_i, h_i]$  and the total weight  $W$  is bounded as  $L \leq W \leq H$ . The locus  $C^\alpha(\mathbf{x}) = C_{1,1}(X)$  if  $[l_i, h_i] = [0, \hat{w}_i]$ .

Now suppose that  $\mathbf{y} = (Y, u) \in \mathbf{D}^{K,n}$  is a lighter distribution than  $\mathbf{x}$ . In the previous section we argued that the EMD is bounded below by the minimum ground distance from  $\bar{\mathbf{y}}$  to a point in  $C^{u_\Sigma/w_\Sigma}(\mathbf{x})$ . We denote this minimum distance as  $\text{CLOC}(\mathbf{x}, \mathbf{y})$  because it uses the locus of the centroid of sub-distributions of  $\mathbf{x}$  of weight  $u_\Sigma$ . Thus  $\text{EMD}(\mathbf{x}, \mathbf{y}) \geq \text{CLOC}(\mathbf{x}, \mathbf{y})$ . If  $d = L_2$ , then this lower bound can be computed by minimizing a quadratic objective function subject to linear constraints:

$$(\text{CLOC}(\mathbf{x}, \mathbf{y}))^2 = \min_v \|Xv \leftrightarrow \bar{\mathbf{y}}\|_2^2$$

subject to

$$\begin{aligned} v &\geq 0 \\ v &\leq \hat{w} = \frac{1}{u_\Sigma} w \\ 1^T v &= 1. \end{aligned}$$

The above minimization problem consists of  $m$  variables and  $2m+1$  linear constraints which are taken directly from (5.10). It can be written more compactly as

$$(\text{CLOC}(\mathbf{x}, \mathbf{y}))^2 = \min_{p \in C^{u_\Sigma/w_\Sigma}(\mathbf{x})} \|p \leftrightarrow \bar{\mathbf{y}}\|_2^2, \quad (5.11)$$

where it is assumed that  $u_\Sigma \leq w_\Sigma$ .

### 5.1.2.2 The Centroid Bounding Box Lower Bound

As previously mentioned, the computation of the CLOC lower bound as described in the previous section is likely to require more time than an exact EMD computation. Yet the centroid locus  $C^\alpha(\mathbf{x})$  can still be very useful in finding a fast to compute lower bound on the EMD. The idea is to precompute a bounding box  $B^\alpha(\mathbf{x})$  for  $C^\alpha(\mathbf{x})$  for a sample of  $\alpha$  values, say  $\alpha = 0.05k$  for  $k = 1, \dots, 20$ . When given a lighter query distribution  $\mathbf{y}$  at query time, the minimum distance from  $\bar{\mathbf{y}}$  to the bounding box  $B^{\alpha_{\mathbf{y}}}(\mathbf{x})$  is a lower bound on  $\text{EMD}(\mathbf{x}, \mathbf{y})$ , where  $\alpha_{\mathbf{y}}$  is the largest sample  $\alpha$  value which does not exceed the total weight ratio  $u_\Sigma/w_\Sigma$  (the correctness of  $\alpha_{\mathbf{y}}$  follows from the containment property (5.14)). We call this lower bound the CBOX lower bound (the  $C$  stands for *centroid* and the  $BOX$  comes from *bounding box*), and it is formally defined as

$$\text{CBOX}(\mathbf{x}, \mathbf{y}) = \min_{p \in B^{u_\Sigma/w_\Sigma}(\mathbf{x})} \|p \leftrightarrow \bar{\mathbf{y}}\|, \quad (5.12)$$

where, once again, it is assumed that  $u_\Sigma \leq w_\Sigma$ . This lower bound computation will be very fast because the bounding boxes are precomputed and the query time computation of the minimum distance of the point  $\bar{\mathbf{y}}$  to the box  $B^{\alpha_{\mathbf{y}}}(\mathbf{x})$  is a constant time operation (it is linear in the dimension  $K$ , but does not depend on the number of points in  $\mathbf{x}$  or  $\mathbf{y}$ ). When  $d = L_2^2$ , we replace the norm  $\|\cdot\|$  in (5.12) with  $\|\cdot\|_2^2$ .

If we write the matrix  $X$  in terms of its rows as

$$X = \begin{bmatrix} r_1^T \\ \vdots \\ r_K^T \end{bmatrix} \in \mathbf{R}^{K \times m}, \quad \text{then} \quad Xv = \begin{bmatrix} r_1^T v \\ \vdots \\ r_K^T v \end{bmatrix} \in \mathbf{R}^K.$$

The computation of an axis-aligned bounding box for the centroid locus  $C^\alpha(x)$  can be accomplished by solving the  $2K$  linear programs

$$a_k = \min_v r_k^T v, \quad b_k = \max_v r_k^T v \quad k = 1, \dots, K$$

subject to

$$\begin{aligned} v &\geq 0 \\ v &\leq \hat{w} = \frac{1}{\alpha w_\Sigma} w \\ \mathbf{1}^T v &= 1. \end{aligned} \quad (5.13)$$

Each of these linear programs has  $m$  variables and  $2m + 1$  constraints. The axis-aligned bounding box for the centroid locus  $C^\alpha(\mathbf{x})$  is

$$B^\alpha(\mathbf{x}) = \prod_{k=1}^K [a_k, b_k].$$

As with the true centroid loci  $C^\alpha(\mathbf{x})$ , we have a containment property for the bounding boxes  $B^\alpha(\mathbf{x})$ :

$$B^{\alpha_1}(\mathbf{x}) \supseteq B^{\alpha_2}(\mathbf{x}) \quad \text{if } \alpha_1 \leq \alpha_2. \quad (5.14)$$

This fact can be verified by observing that the constraints over which the minima  $a_k$  and maxima  $b_k$  are computed get weaker as  $\alpha$  decreases (the only constraint involving  $\alpha$  is (5.13)). Note also that the box  $B^\alpha(\mathbf{x})$  includes its “interior” so that the lower bound  $\text{CBOX}(\mathbf{x}, \mathbf{y})$  is zero if  $\bar{\mathbf{y}}$  lies inside  $B^{\alpha\mathbf{y}}(\mathbf{x})$ . Using the CBOX lower bound instead of the CLOC lower bound trades off computation speed for pruning power since the former is much faster to compute, but<sup>1</sup>

$$\text{EMD}(\mathbf{x}, \mathbf{y}) \geq \text{CLOC}(\mathbf{x}, \mathbf{y}) \geq \text{CBOX}(\mathbf{x}, \mathbf{y}).$$

Nevertheless, the pruning power of the CBOX lower bound will be high when the query distribution is well-separated from many of the database distributions (which implies that the centroids will also be well-separated).

## 5.2 Projection-based Lower Bounds

For  $v$  on the unit sphere  $S^{K-1}$  in  $\mathbf{R}^K$ , the projection  $\text{proj}_v(\mathbf{x})$  of the distribution  $\mathbf{x} = (X, w) \in \mathbf{R}^{K,m}$  along the direction  $v$  is defined as

$$\text{proj}_v(\mathbf{x}) = \{ (v^T x_1, w_1), (v^T x_2, w_2), \dots, (v^T x_m, w_m) \} = (v^T X, w) \in \mathbf{D}^{1,m}.$$

In words, the projection along  $v$  is obtained by using the lengths of the projections of the distribution points along  $v$  and leaving the corresponding weights unchanged. The following lemma shows that the EMD between projections is a lower bound on the EMD between the original distributions. See Figure 5.1.

**Lemma 3** *Let  $v \in S^{K-1}$ . Then  $\text{EMD}^{L_2}(\mathbf{x}, \mathbf{y}) \geq \text{EMD}^{L_1}(\text{proj}_v(\mathbf{x}), \text{proj}_v(\mathbf{y}))$ .*

---

<sup>1</sup>The inequality  $\text{CLOC}(\mathbf{x}, \mathbf{y}) \geq \text{CBOX}(\mathbf{x}, \mathbf{y})$  follows from the fact that  $B^\alpha(\mathbf{x}) \supseteq C^\alpha(\mathbf{x})$  (since  $B^\alpha(\mathbf{x})$  is a bounding box for  $C^\alpha(\mathbf{x})$ ) and the definitions (5.11) and (5.12).



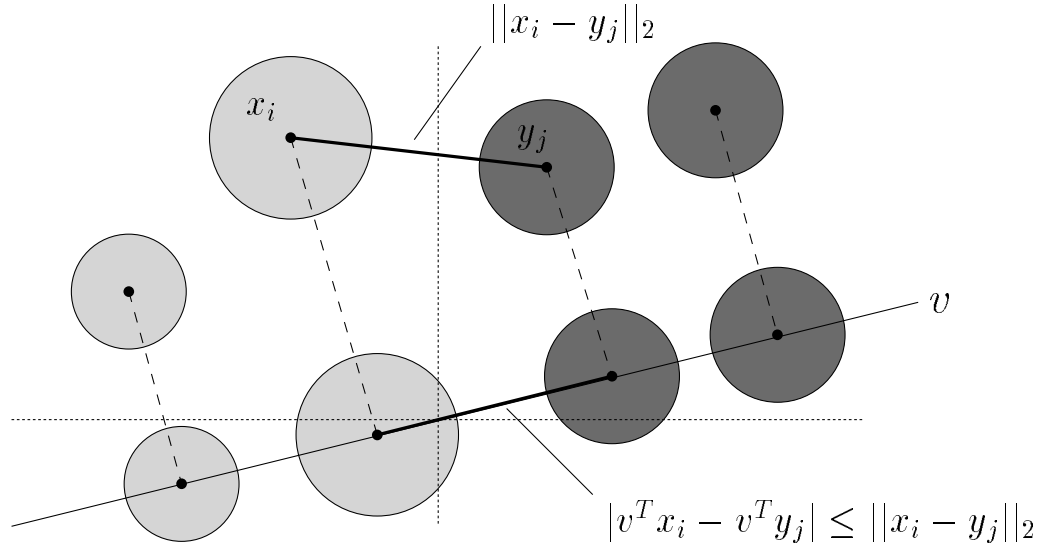


Figure 5.1: The Projection Lower Bound. The EMD with  $d = L_2$  between two distributions is greater than or equal to the EMD with  $d = L_1$  between the projections of the distributions onto a line through the origin. This is because all ground distances decrease or remain the same after projection. See Lemma 3 and its proof.

**Proof.** This theorem follows easily from the definition of the EMD and the fact that

$$\begin{aligned}
 |v^T x_i \leftrightarrow v^T y_j| &= |v^T(x_i \leftrightarrow y_j)| \\
 &= \|v\|_2 \|x_i \leftrightarrow y_j\|_2 |\cos \theta_{v, (x_i - y_j)}| \\
 &= \|x_i \leftrightarrow y_j\|_2 |\cos \theta_{v, (x_i - y_j)}| \\
 |v^T x_i \leftrightarrow v^T y_j| &\leq \|x_i \leftrightarrow y_j\|_2.
 \end{aligned}$$

■

The following theorem is an immediate consequence of Lemma 3.

**Theorem 8** Let  $V = \{v_1, \dots, v_L\} \subset S^{K-1}$  and

$$\text{PMAX}(V, \mathbf{x}, \mathbf{y}) = \max_{v \in V} \text{EMD}(\text{proj}_v(\mathbf{x}), \text{proj}_v(\mathbf{y}))$$

Then  $\text{EMD}(\mathbf{x}, \mathbf{y}) \geq \text{PMAX}(V, \mathbf{x}, \mathbf{y})$ .

For this lower bound to be of practical use, we must be able to compute it efficiently. In section 4.3.2, we presented a straightforward,  $\Theta(m + n)$  time algorithm to compute the EMD between equal-weight distributions on the real line. In combination with Theorem 8, this algorithm provides the means to compute quickly a lower bound on the EMD between

two equal-weight distributions.

One pruning strategy is to pick a set of random directions  $V$  along which to perform projections, and apply Theorem 8 to obtain a lower bound. The hope is that the differences between two distributions will be captured by looking along one of the directions in  $V$ . Another pruning strategy is to use the set of orthogonal axis directions for the set  $V$ . The following corollary is an immediate consequence of Theorem 8.

**Corollary 3** *Let  $E = \{e_1, \dots, e_K\} \subset S^{K-1}$  be the set of axis directions, and let*

$$\text{PAMAX}(\mathbf{x}, \mathbf{y}) = \text{PMAX}(E, \mathbf{x}, \mathbf{y}).$$

*Then  $\text{EMD}(\mathbf{x}, \mathbf{y}) \geq \text{PAMAX}(\mathbf{x}, \mathbf{y})$ .*

Looking along the space axes is intuitively appealing when each axis measures a specific property. For example, suppose that distribution points are points in the CIE-Lab color space ([88]). If two images are very different in terms of the luminance values of pixels, then comparing the signature projections along the L-axis will reveal this difference and allow the system to avoid an exact EMD computation.

When the projection directions are the coordinate axes, we can prove a lower bound which involves the sum of the EMDs along axis directions.

**Theorem 9** *If*

$$\text{PASUM}(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{K}} \sum_{k=1}^K \text{EMD}(\text{proj}_{e_k}(\mathbf{x}), \text{proj}_{e_k}(\mathbf{y})),$$

*then  $\text{EMD}(\mathbf{x}, \mathbf{y}) \geq \text{PASUM}(\mathbf{x}, \mathbf{y})$ .*

**Proof.** The proof uses the fact that  $\|a\|_2 \geq (1/\sqrt{K})\|a\|_1$  for any vector  $a \in \mathbf{R}^K$  ([25]). It follows that

$$\begin{aligned} \sum_{i=1}^m \sum_{j=1}^n f_{ij} \|x_i \Leftrightarrow y_j\|_2 &\geq \frac{1}{\sqrt{K}} \sum_{i=1}^m \sum_{j=1}^n f_{ij} \|x_i \Leftrightarrow y_j\|_1 \\ &= \frac{1}{\sqrt{K}} \sum_{i=1}^m \sum_{j=1}^n f_{ij} \sum_{k=1}^K |x_i^{(k)} \Leftrightarrow y_j^{(k)}| \\ &= \frac{1}{\sqrt{K}} \sum_{k=1}^K \sum_{i=1}^m \sum_{j=1}^n f_{ij} |x_i^{(k)} \Leftrightarrow y_j^{(k)}| \\ \sum_{i=1}^m \sum_{j=1}^n f_{ij} \|x_i \Leftrightarrow y_j\|_2 &\geq \frac{1}{\sqrt{K}} \sum_{k=1}^K \sum_{i=1}^m \sum_{j=1}^n f_{ij} |x_i^{(k)} \Leftrightarrow y_j^{(k)}|, \end{aligned}$$

where the superscript  $(k)$  denotes the  $k$ th component of a vector. Therefore,

$$\begin{aligned}
\min_{F \in \mathcal{F}(\mathbf{x}, \mathbf{y})} \sum_{i=1}^m \sum_{j=1}^n f_{ij} \|x_i \Leftrightarrow y_j\|_2 &\geq \min_{F \in \mathcal{F}(\mathbf{x}, \mathbf{y})} \frac{1}{\sqrt{K}} \sum_{k=1}^K \sum_{i=1}^m \sum_{j=1}^n f_{ij} |x_i^{(k)} \Leftrightarrow y_j^{(k)}| \\
&\geq \frac{1}{\sqrt{K}} \sum_{k=1}^K \min_{F \in \mathcal{F}(\mathbf{x}, \mathbf{y})} \sum_{i=1}^m \sum_{j=1}^n f_{ij} |x_i^{(k)} \Leftrightarrow y_j^{(k)}| \\
&= \frac{1}{\sqrt{K}} \sum_{k=1}^K (\min(w_\Sigma, u_\Sigma) \times \text{EMD}(\text{proj}_{e_k}(\mathbf{x}), \text{proj}_{e_k}(\mathbf{y}))) \\
&= \frac{1}{\sqrt{K}} \min(w_\Sigma, u_\Sigma) \sum_{k=1}^K \text{EMD}(\text{proj}_{e_k}(\mathbf{x}), \text{proj}_{e_k}(\mathbf{y})) \\
\min_{F \in \mathcal{F}(\mathbf{x}, \mathbf{y})} \sum_{i=1}^m \sum_{j=1}^n f_{ij} \|x_i \Leftrightarrow y_j\|_2 &\geq \frac{1}{\sqrt{K}} \min(w_\Sigma, u_\Sigma) \sum_{k=1}^K \text{EMD}(\text{proj}_{e_k}(\mathbf{x}), \text{proj}_{e_k}(\mathbf{y})).
\end{aligned}$$

Dividing both sides of the last inequality by  $\min(w_\Sigma, u_\Sigma)$  gives the desired result.  $\blacksquare$

Note that  $\text{PASUM}(\mathbf{x}, \mathbf{y})$  may be rewritten as

$$\text{PASUM}(\mathbf{x}, \mathbf{y}) = \sqrt{K} \left( \frac{\sum_{k=1}^K \text{EMD}(\text{proj}_{e_k}(\mathbf{x}), \text{proj}_{e_k}(\mathbf{y}))}{K} \right).$$

This alternate expression makes it clear that  $\text{PASUM}(\mathbf{x}, \mathbf{y})$  is a better lower bound than  $\text{PAMAX}(\mathbf{x}, \mathbf{y})$  iff the square root of the dimension times the average axis projection distance is greater than the maximum axis projection distance.

Our projection bounds require EMD computations between distributions on the real line. In section 4.3.2, we gave a very efficient algorithm to compute the EMD between equal-weight distributions (with the  $L_1$ -distance as the ground distance). If the distributions have different total weight, we must fall back on the transportation simplex method to compute the 1D EMD. Using arguments similar to those used in section 4.3.2, we can, however, compute a lower bound on the EMD between unequal-weight distributions on the line. The idea to determine intervals over which certain amounts of mass must flow in any feasible flow.

Once again consider the interval  $(r_k, r_{k+1})$ , and WLOG assume  $w_\Sigma > u_\Sigma$  and that  $\mathbf{x}$ -weight is moved to match all the  $\mathbf{y}$ -weight. When there is more  $\mathbf{x}$ -weight than  $\mathbf{y}$ -weight in both  $(\Leftrightarrow\infty, r_k]$  and  $[r_{k+1}, \infty)$ , then there will be feasible flows in which no  $\mathbf{x}$ -weight travels through  $(r_k, r_{k+1})$ . If there is more  $\mathbf{x}$ -weight than  $\mathbf{y}$ -weight in  $(\Leftrightarrow\infty, r_k]$ , but less  $\mathbf{x}$ -weight than  $\mathbf{y}$ -weight in  $[r_{k+1}, \infty)$ , then  $(u_\Sigma \Leftrightarrow U(r_k)) \Leftrightarrow (w_\Sigma \Leftrightarrow W(r_k))$  of the  $\mathbf{x}$ -weight must be moved from  $r_k$  to  $r_{k+1}$  in order to cover the  $\mathbf{y}$ -weight in  $[r_{k+1}, \infty)$ . See Figure 5.2(a). If there is less  $\mathbf{x}$ -weight than  $\mathbf{y}$ -weight in  $(\Leftrightarrow\infty, r_k]$ , but more  $\mathbf{x}$ -weight than  $\mathbf{y}$ -weight in  $[r_{k+1}, \infty)$ ,

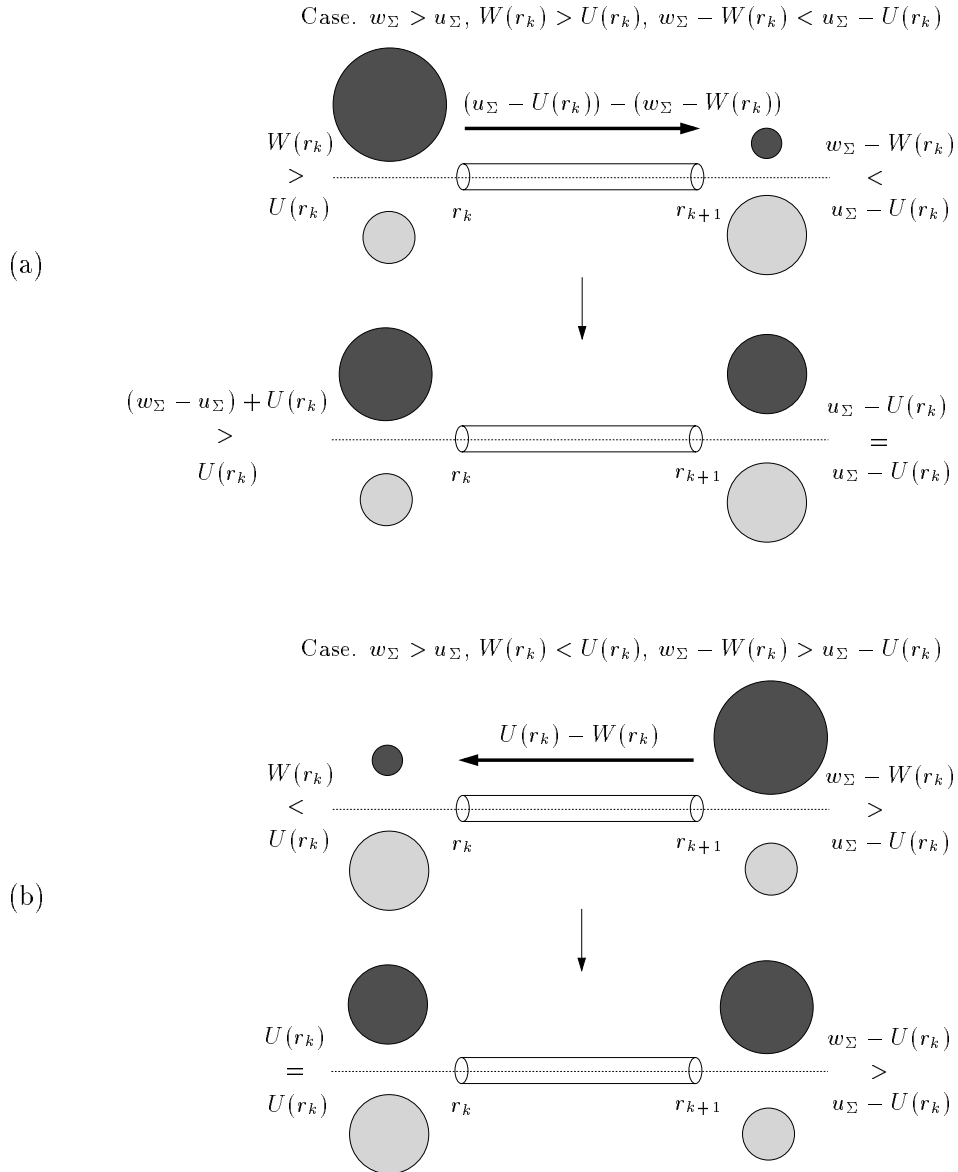


Figure 5.2: Flow Feasibility for Unequal-Weight Distributions on the Real Line.  $\mathbf{x} = (X, w)$  and  $\mathbf{y} = (Y, u)$  are distributions in 1D with  $w_\Sigma > u_\Sigma$ . All  $\mathbf{y}$ -weight must be covered by  $\mathbf{x}$ -weight. (a)  $W(r_k) > U(r_k), w_\Sigma \Leftrightarrow W(r_k) < u_\Sigma \Leftrightarrow U(r_k)$ . In any feasible flow from  $\mathbf{x}$  to  $\mathbf{y}$ , at least  $(w_\Sigma \Leftrightarrow W(r_k)) \Leftrightarrow (u_\Sigma \Leftrightarrow U(r_k))$  of  $\mathbf{x}$ -weight must travel from  $r_k$  to  $r_{k+1}$  during the flow. (b)  $W(r_k) < U(r_k), w_\Sigma \Leftrightarrow W(r_k) > u_\Sigma \Leftrightarrow U(r_k)$ . In any feasible flow from  $\mathbf{x}$  to  $\mathbf{y}$ , at least  $U(r_k) \Leftrightarrow W(r_k)$  of  $\mathbf{x}$ -weight must travel from  $r_{k+1}$  to  $r_k$  during the flow.

then  $U(r_k) \Leftrightarrow W(r_k)$  of the  $\mathbf{x}$ -weight must be moved from  $r_{k+1}$  to  $r_k$  in order to cover the  $\mathbf{y}$ -weight in  $(\Leftrightarrow\infty, r_k]$ . This case is illustrated in Figure 5.2(b). Under the assumption that  $w_\Sigma > u_\Sigma$ , it *cannot* be the case that there is less  $\mathbf{x}$ -weight than  $\mathbf{y}$ -weight in both  $(\Leftrightarrow\infty, r_k]$  and  $[r_{k+1}, \infty)$ .

Pseudocode for the lower bound described in the previous paragraph is given below. The routine is named FSBL because the lower bound follows simply from flow feasibility (FeaSiBiLity) conditions.

```

function emdlb = FSBL( $\mathbf{x}, \mathbf{y}$ )
/* assumes  $K = 1$ ,  $w_\Sigma \geq u_\Sigma$ , ground distance is  $L_1$  */
/* assumes  $x_1 \leq x_2 \leq \dots \leq x_m$ ,  $y_1 \leq y_2 \leq \dots \leq y_n$  */
  work = wcumsum = ucumsum = r = 0
  /* first increment of work will be 0, regardless of  $r$  */
  wsum =  $\sum_{i=1}^m w_i$ 
  usum =  $\sum_{j=1}^n u_j$ 
  i = j = 1
  xnext =  $x_1$ 
  ynext =  $y_1$ 
  while (( $i \leq m$ ) or ( $j \leq n$ ))
    next = min(xnext, ynext)
    if (usum - ucumsum > wsum - wcumsum)
      work += ((usum - ucumsum) - (wsum - wcumsum)) * (next - r)
    elseif (ucumsum > wcumsum)
      work += (ucumsum - wcumsum) * (next - r)
    end if
    if (xnext  $\leq$  ynext)
      wcumsum +=  $w_i$ 
      i += 1
      xnext = ( $i \leq m$ ) ?  $x_i$  :  $\infty$ 
    else
      ucumsum +=  $u_j$ 
      j += 1
      ynext = ( $j \leq n$ ) ?  $y_j$  :  $\infty$ 
    end if
    r = next
  end while
  return (work / usum)
end function

```

We have argued that

**Theorem 10** *If  $\mathbf{x}$  and  $\mathbf{y}$  are distributions on the real line, then  $\text{EMD}(\mathbf{x}, \mathbf{y}) \geq \text{FSBL}(\mathbf{x}, \mathbf{y})$ .*

If  $w_\Sigma = u_\Sigma$ , then  $(u_\Sigma \Leftrightarrow U(r_k) > w_\Sigma \Leftrightarrow W(r_k)) \equiv (W(r_k) > U(r_k))$ ,  $(u_\Sigma \Leftrightarrow U(r_k)) \Leftrightarrow (w_\Sigma \Leftrightarrow W(r_k)) = W(r_k) \Leftrightarrow U(r_k)$ , and the routine computes the exact value  $\text{EMD}(\mathbf{x}, \mathbf{y})$ .

**Theorem 11** *If  $\mathbf{x}$  and  $\mathbf{y}$  are equal-weight distributions on the real line, then  $\text{EMD}(\mathbf{x}, \mathbf{y}) = \text{FSBL}(\mathbf{x}, \mathbf{y})$ .*

Assuming that the points in  $\mathbf{x} \in \mathbf{D}^{1,m}$  and  $\mathbf{y} \in \mathbf{D}^{1,n}$  are in sorted order, the routine FSBL runs in linear time  $\Theta(m+n)$ . The combined sorted list  $r_1, \dots, r_{m+n}$  of points in  $\mathbf{x}$  and  $\mathbf{y}$  is discovered by walking along the two sorted lists of points. At any time during the algorithm, there is a pointer to the next  $\mathbf{x}$  and next  $\mathbf{y}$  value to be considered. The value  $r_{k+1}$  then follows in constant time from the value of  $r_k$ .

The FSBL lower bound may be substituted for the EMD function in the PMAX, PAMAX, and PASUM lower bounds to obtain efficient to compute, projection-based lower bounds

$$\begin{aligned} \text{PMAX}_{\text{FSBL}}(V, \mathbf{x}, \mathbf{y}) &= \max_{v \in V} \text{FSBL}(\text{proj}_v(\mathbf{x}), \text{proj}_v(\mathbf{y})) \\ &= \text{PMAX}(V, \mathbf{x}, \mathbf{y}) \quad \text{when } w_\Sigma = u_\Sigma \end{aligned}$$

$$\begin{aligned} \text{PAMAX}_{\text{FSBL}}(\mathbf{x}, \mathbf{y}) &= \max_{k=1, \dots, K} \text{FSBL}(\text{proj}_{e_k}(\mathbf{x}), \text{proj}_{e_k}(\mathbf{y})) \\ &= \text{PAMAX}(\mathbf{x}, \mathbf{y}) \quad \text{when } w_\Sigma = u_\Sigma \end{aligned}$$

$$\begin{aligned} \text{PASUM}_{\text{FSBL}}(\mathbf{x}, \mathbf{y}) &= \frac{1}{\sqrt{K}} \sum_{k=1}^K \text{FSBL}(\text{proj}_{e_k}(\mathbf{x}), \text{proj}_{e_k}(\mathbf{y})) \\ &= \text{PASUM}(\mathbf{x}, \mathbf{y}) \quad \text{when } w_\Sigma = u_\Sigma \end{aligned}$$

in which  $\mathbf{x}$  and  $\mathbf{y}$  are not necessarily equal weight. The second equality in each of the three pairs of equalities follows directly from Theorem 11 and the definitions of  $\text{PMAX}(V, \mathbf{x}, \mathbf{y})$ ,  $\text{PAMAX}(\mathbf{x}, \mathbf{y})$ , and  $\text{PASUM}(\mathbf{x}, \mathbf{y})$ .

### 5.3 Experiments in Color-based Retrieval

In this section, we show some results of using the lower bounds  $\text{CBOX}$ ,  $\text{PMAX}_{\text{FSBL}}$ ,  $\text{PAMAX}_{\text{FSBL}}$ , and  $\text{PASUM}_{\text{FSBL}}$  in the color-based retrieval system described in [65]. This system summarizes an image by a distribution of dominant colors in the CIE-Lab color space, where the weight of a dominant color is equal to the fraction of image pixels which are classified as that color. The input to the system is a query and a number  $R$  of nearest

images to return. The system computes the EMD between the query distribution and each of the database distributions. The ground distance is  $d = L_2$ .

If the query is a full image (e.g. an image in the database), then the query distribution and all the database distributions will have total weight equal to one. In this query-by-example setting, the system first checks the distance between distribution centroids before performing an exact EMD computation. If the centroid distance is larger than the  $R$ th smallest distance seen before the current comparison, then the system does not compute the EMD and simply considers the next database image. An  $R$ -nearest neighbor database image to the query cannot be missed by this algorithm because the centroid distance is a lower bound on the EMD between equal-weight distributions. When the query is a partial query (such as “give me all the images with at least 20% sky blue”), the system in [65] performs an exact EMD computation between the query and every database image.

To use the CBOX lower bound for partial queries, some additional preprocessing is needed. At database entry time, the distribution  $\mathbf{x} = (X, w)$  of an image is computed and stored, as well as the centroid bounding boxes  $B^\alpha(\mathbf{x})$  for  $\alpha = 0.05k, k = 1, \dots, 20$ . Given a query distribution  $\mathbf{y} = (Y, u)$  of weight  $u_\Sigma \leq w_\Sigma$ , let  $\alpha_{\mathbf{y}}$  denote the largest sample  $\alpha$  value which does not exceed the total weight ratio  $u_\Sigma/w_\Sigma$ . The system computes the distance between  $\bar{\mathbf{y}}$  and the nearest point in  $B^{\alpha_{\mathbf{y}}}(\mathbf{x})$ . This is the CBOX lower bound. To use the  $\text{PMAX}_{\text{FSBL}}$  lower bound, a set  $V$  of  $L$  (specified later) random projection directions and the  $L$  position-sorted projections of each database distribution along the directions in  $V$  are computed and stored at database load time. At query time, the query distribution is also projected along the directions in  $V$ . To use the  $\text{PAMAX}_{\text{FSBL}}$  and  $\text{PASUM}_{\text{FSBL}}$  lower bounds, the  $K$  position-sorted projections of each database distribution along the space axes are computed and stored at database entry time. At query time, the same axis projections are performed on the query distribution.

There are many factors that affect the performance of our lower bounds. The most obvious is the database itself. Here, we use a Corel database of 20000 color images which is dominated by outdoor scenes. The order in which the images are compared to the query is also important. If the most similar images to a query are processed first, then the  $R$ th smallest distance seen will be relatively small when the dissimilar images are processed, and relatively weak lower bounds can prune these dissimilar images. Of course, the purpose of the query is to discover the similar images. Nonetheless, a random order of comparison may help ensure good performance over a wide range of queries. Moreover, if a certain type of query is more likely than others, say, for example, queries with large amounts of blue and green (to retrieve outdoor images containing sky and grass), then it would be wise to pre-determine a good comparison order to use for such queries. In the results that follow,

however, the comparison order is the same for all queries, and the order is *not* specialized for any particular type of query.

The number  $R$  of nearest images to return is yet another factor. For a fixed comparison order and query, the number of exact EMD calculations pruned is inversely related to the size of  $R$ . This is because the  $R$ th smallest distance (against which a lower bound is compared) after comparing a fixed number images is an increasing function of  $R$ . In all the upcoming experiments, the number of nearest images returned is fixed at  $R = 20$ . In terms of the actual lower bounds, a system may be able to achieve better query times by using more than one bound. For example, a system might apply the CBOX lower bound first, followed by the more expensive  $\text{PASUM}_{\text{FSBL}}$  bound if CBOX fails, followed by an even more expensive exact EMD computation if  $\text{PASUM}_{\text{FSBL}}$  also fails. The hope is that the lower bound hierarchy of CBOX,  $\text{PASUM}_{\text{FSBL}}$ , and EMD speeds up query times in much the same way that the memory hierarchy of primary cache, secondary cache, and main memory speeds up memory accesses. Our experiments, however, apply one lower bound per query. For the  $\text{PMAX}_{\text{FSBL}}$  lower bound, the number  $L$  of random directions must be specified. This parameter trades off between pruning power and computation speed. The more directions, the greater the pruning power, but the slower the computation. In our work, we use the heuristic  $L = 2K$  (without quantifiable justification), where  $K$  is the dimension of the underlying point space (so  $L = 6$  in the color-based system).

All experiments were conducted on an SGI Indigo<sup>2</sup> with a 250 MHz processor, and query times are reported in seconds (s). The exact EMD is computed by the transportation simplex method as described by Hillier and Lieberman in [32]. The color signature of a typical database image has eight to twelve points. The time for an EMD calculation between two such images varies roughly between half a millisecond and one millisecond (ms). The EMD computation time increases with the number of points in the distributions, so EMD computations involving a partial query distribution with only a few points are, in general, faster than EMD computations between two database images. The time for an EMD computation between a database image and a partial query with three or fewer points is typically about 0.25ms.

We begin our experiments with a few very simple queries. Each of these queries consists of a distribution with exactly one color point in CIE-Lab space. The results of the three queries



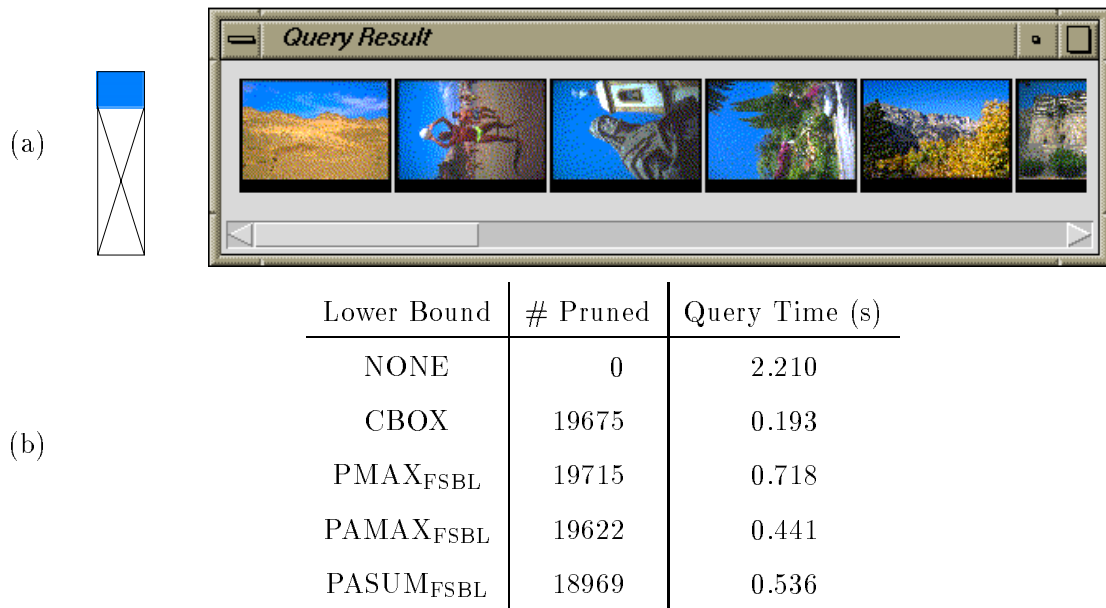
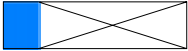
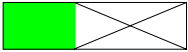



Figure 5.3: Query C.1.1 – 20% Blue. (a) query results. (b) query statistics.

C.1.1	at least 20% (sky) blue		,
C.1.2	at least 40% green		, and
C.1.3	at least 60% red		

are shown in Figure 5.3, Figure 5.4, and Figure 5.5, respectively. In these examples, all the lower bounds result in query times which are less than the brute force query time, and avoid a large fraction of exact EMD computations. The CBOX and  $\text{P}_{\text{ASUM}}_{\text{FSBL}}$  bounds gave the best results on these three queries.

The next set of examples consists of randomly generated partial queries. The results for the five queries

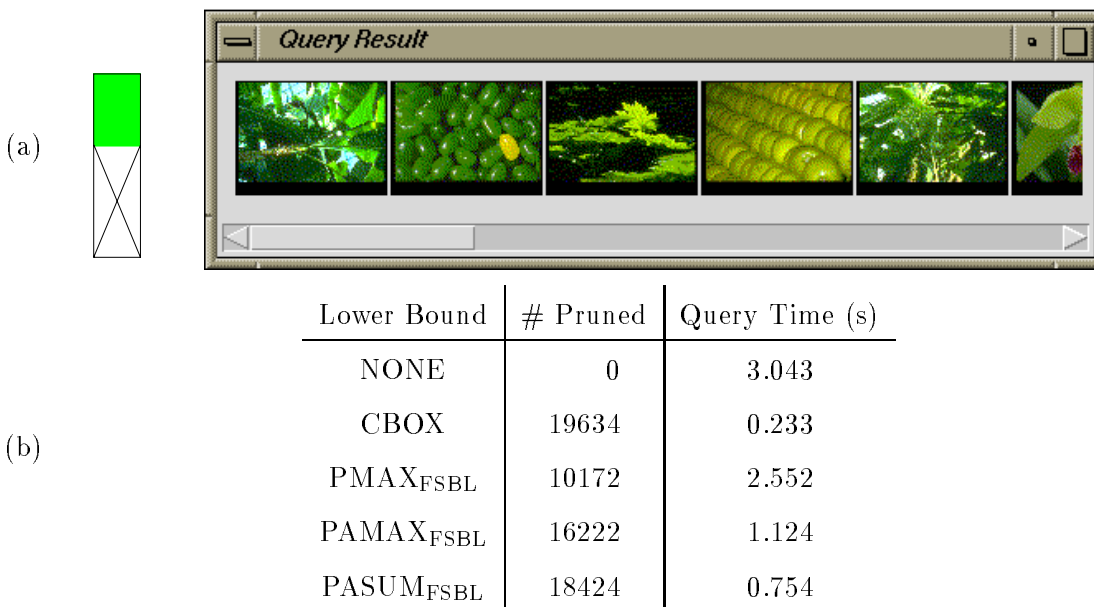


Figure 5.4: Query C.1.2 – 40% Green. (a) query results. (b) query statistics.

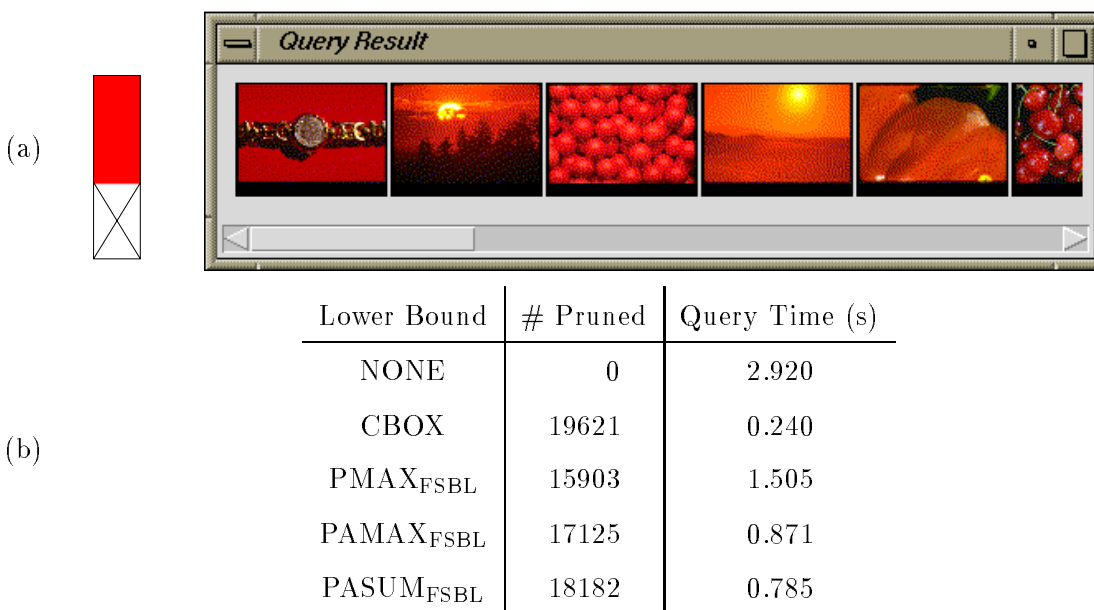


Figure 5.5: Query C.1.3 – 60% Red. (a) query results. (b) query statistics.

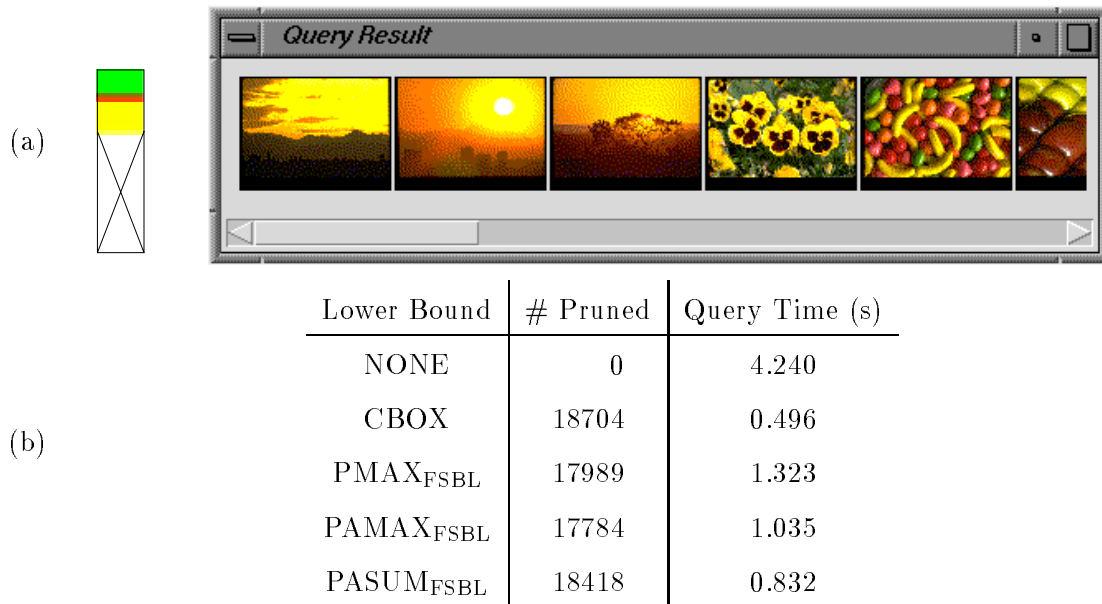


Figure 5.6: Query C.2.1 – 13.5% Green, 3.4% Red, 17.8% Yellow. The total weight of the query is  $u_{\Sigma} = 34.7\%$ . (a) query results. (b) query statistics.

C.2.1	13.5% green, 3.4% red, 17.8% yellow		,
C.2.2	26.0% blue, 19.7% violet		,
C.2.3	16.8% blue, 22.2% green, 1.8% yellow		,
C.2.4	22.8% red, 24.2% green, 17.3% blue		, and
C.2.5	13.2% yellow, 15.3% violet, 15.3% green		

are shown in Figure 5.6 through Figure 5.10, respectively. The CBOX lower bound gives the best results for queries C.2.1 and C.2.2, but its performance drops by an order of magnitude for C.2.3, and it is completely ineffective for C.2.4 and C.2.5. Indeed, the CBOX lower bound pruned only 1 of 20000 database images for query C.2.5. The CBOX behavior can be explained in part by the locations of centroids of the query distributions and the database distributions. See Figure 5.11. Roughly speaking, the effectiveness of the CBOX bound is directly related to the amount of separation between the database distributions and the query distribution, with larger separation implying a more effective bound. The

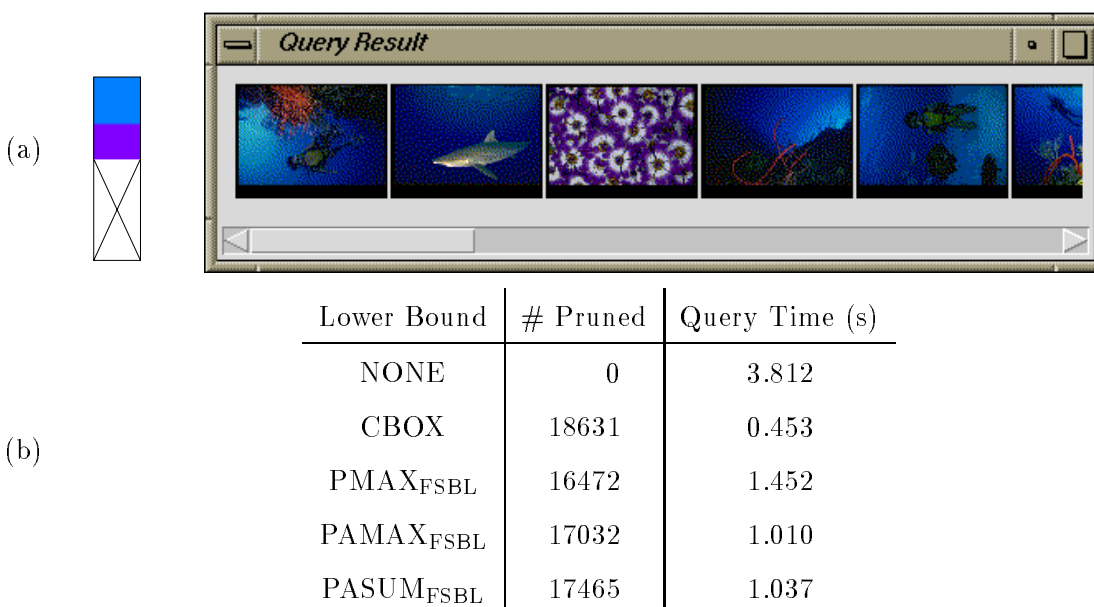


Figure 5.7: Query C.2.2 – 26.0% Blue, 19.7% Violet. The total weight of the query is  $u_{\Sigma} = 45.7\%$ . (a) query results. (b) query statistics.

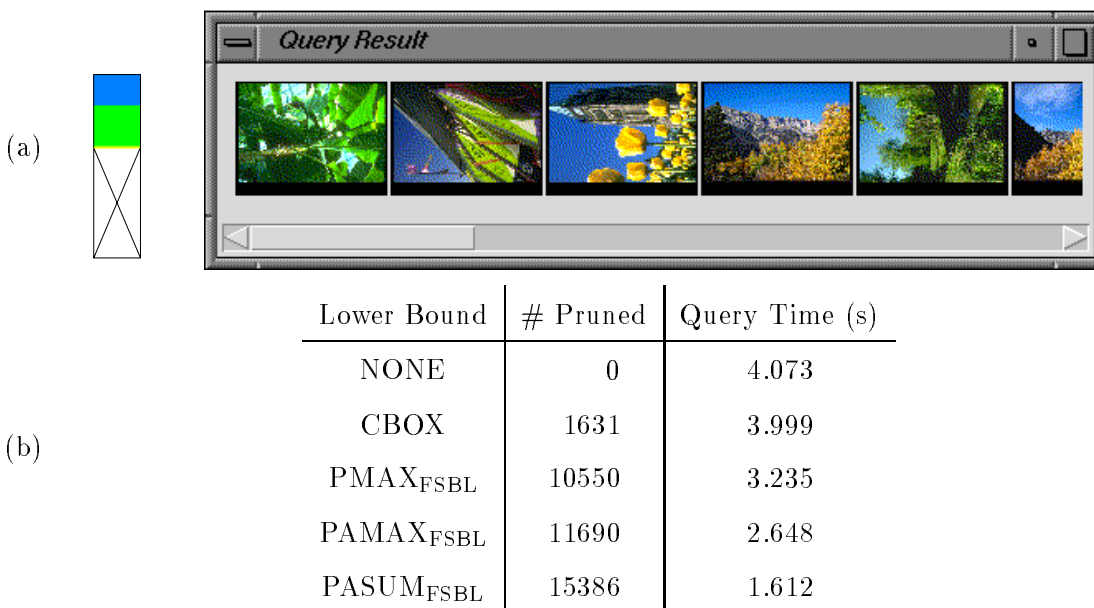


Figure 5.8: Query C.2.3 – 16.8% Blue, 22.2% Green, 1.8% Yellow. The total weight of the query is  $u_{\Sigma} = 40.8\%$ . (a) query results. (b) query statistics.

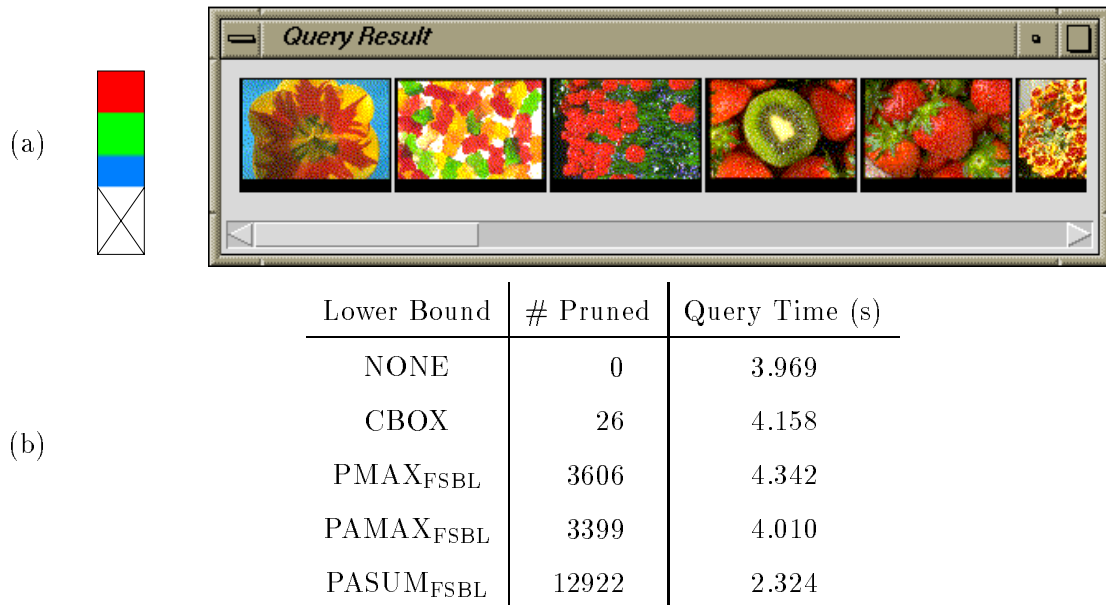


Figure 5.9: Query C.2.4 – 22.8% Red, 24.2% Green, 17.3% Blue. The total weight of the query is  $u_{\Sigma} = 64.3\%$ . (a) query results. (b) query statistics.

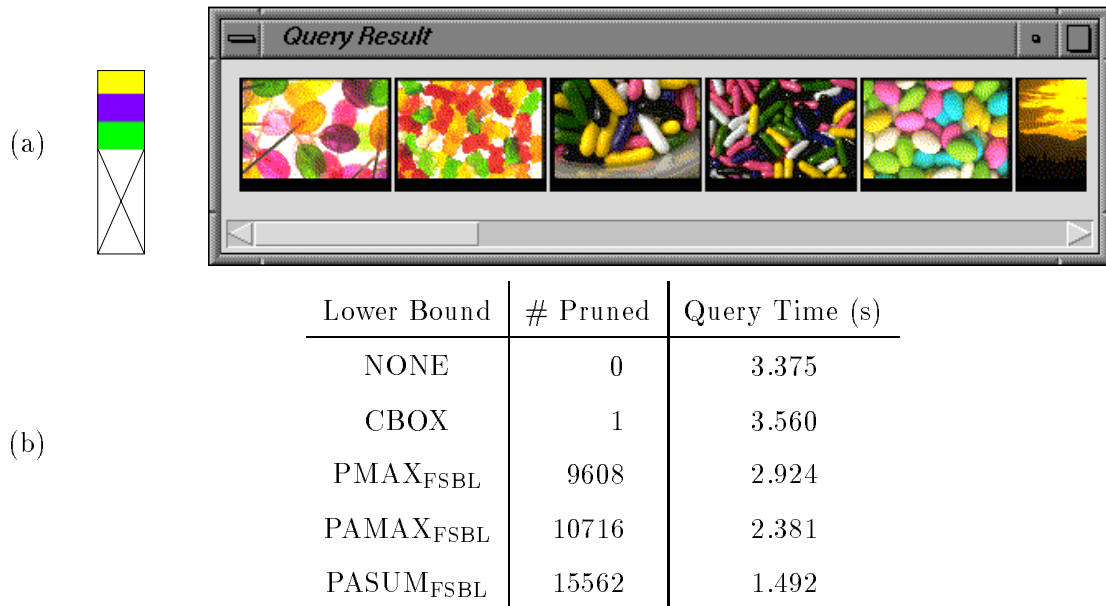


Figure 5.10: Query C.2.5 – 13.2% Yellow, 15.3% Violet, 15.3% Green. The total weight of the query is  $u_{\Sigma} = 43.8\%$ . (a) query results. (b) query statistics.

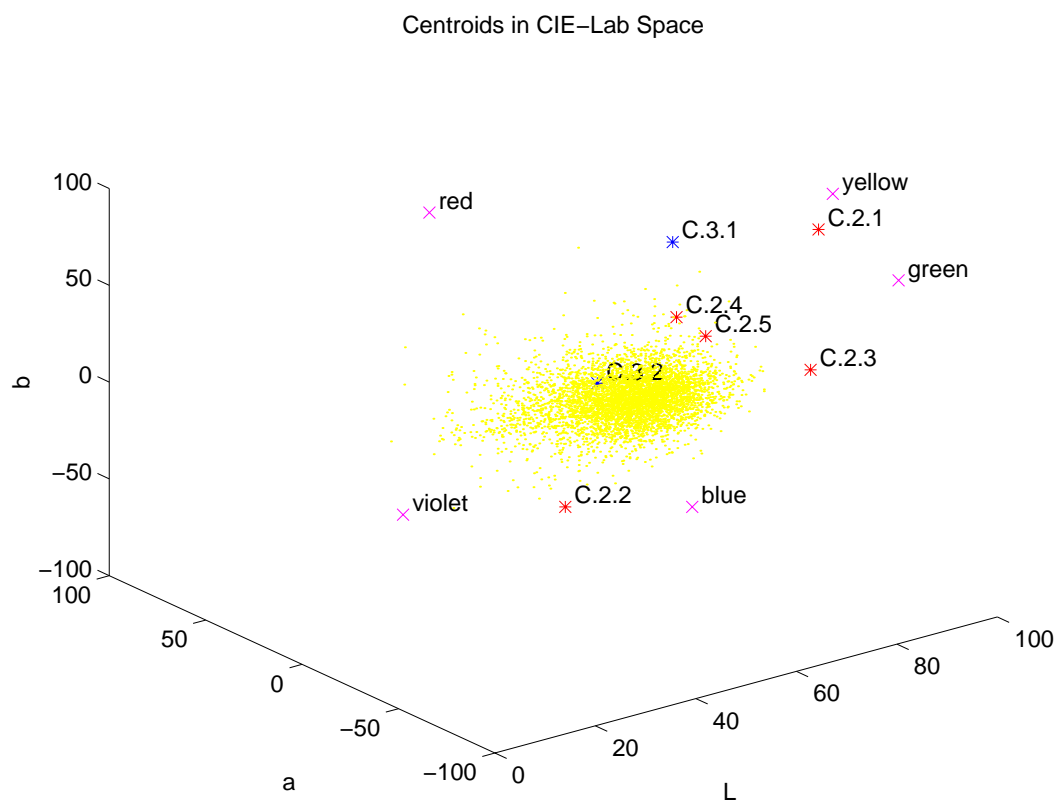


Figure 5.11: Distribution Centroids for Corel Database Images and Example Queries. The centroids of the color signature distributions of a random subset of 5000 images in the Corel database are plotted as dots, and the centroids for the queries C.2.\* and C.3.\* are plotted as stars. The locations of blue (C.1.1), green (C.1.2), red (C.1.3), yellow, and violet are plotted as x's.

query C.2.1 consists almost entirely of green and yellow. As one can see from Figure 5.11, the centroid of C.2.1 is very isolated from the database centroids. The approximately equal amounts red, green, and blue in query C.2.4 result in a centroid which is close to a large number of database centroids. The same statement holds for query C.2.5 which has green and yellow in one corner of the CIE-Lab space, and violet at the opposite corner.

The distances of the centroids for C.2.2 and C.2.3 to the database centroids are (i) about the same, and (ii) are smaller than the distance for C.2.1 and larger than the distances for C.2.4 and C.2.5. Observation (ii) helps explain why the performance of CBOX on C.2.2 and C.2.3 is worse than the performance on C.2.1, but better than the performance on C.2.4 and C.2.5. Observation (i) might lead one to believe that the CBOX performance should be about the same on C.2.2 and C.2.3. The statistics, however, show that this is not the case. To understand why, we must remember that the queries are partial queries. The relevant quantity is not the centroid of a database distribution, but rather the locus of the centroid of all sub-distributions with weight equal to the weight of the query. Consider images with significant amounts of blue and green, and other colors which are distant from blue and green (such as red). The other colors will help move the distribution centroid away from blue and green. However, a sub-distribution of such an image which contains only blue and green components will have a centroid which is close to blue and green, and hence close to the centroid of C.2.3. The distance between the query centroid and this image centroid may be large, but the CBOX lower bound will be small (and, hence, weak). From Figure 5.11 and the results of C.2.2 and C.2.3, one can infer that there are many more images that contain blue, green, and significant amounts of distant colors from blue and green than there are images that contain blue, violet, and significant amounts of distant colors from blue and violet. The centroid is a measure of the (weighted) average color in a distribution, and the average is not an accurate representative of a distribution with high variance (i.e. with colors that span a large portion of the color space).

The projection-based lower bounds  $\text{PMAX}_{\text{FSBL}}$ ,  $\text{PAMAX}_{\text{FSBL}}$ ,  $\text{PASUM}_{\text{FSBL}}$  compare two distributions by comparing the distributions projected along some set of directions. There is hope that these bounds will help when the CBOX bound is ineffective. In queries C.2.3, C.2.4, and C.2.5, the projection-based lower bounds prune far more EMD calculations than the CBOX bound. However, pruning a large number of EMD calculations does *not* guarantee a smaller query time than achievable by brute force because of the overhead of computing a lower bound when it fails to prune an EMD calculation. In all the random partial queries C.2.\*, the query times for  $\text{PMAX}_{\text{FSBL}}$ ,  $\text{PAMAX}_{\text{FSBL}}$ , and  $\text{PASUM}_{\text{FSBL}}$  were less than the query times for brute force processing, except for the  $\text{PMAX}_{\text{FSBL}}$  and  $\text{PAMAX}_{\text{FSBL}}$  bounds in query C.2.4. In particular, the  $\text{PASUM}_{\text{FSBL}}$  bound performed very

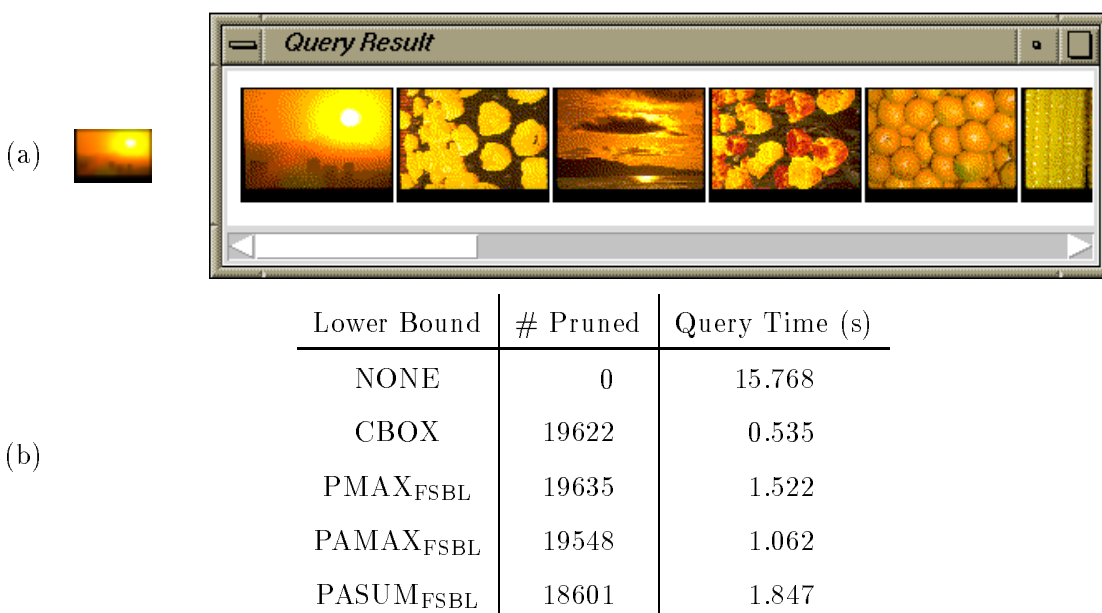
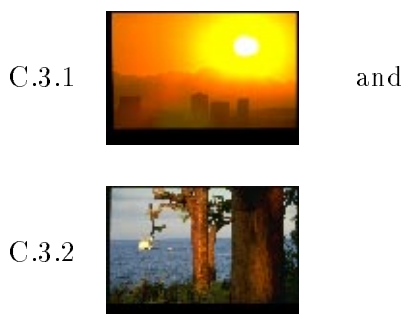


Figure 5.12: Query C.3.1 – Sunset Image. (a) query results. (b) query statistics.

well for all the queries. Since the projection-based lower bounds are more expensive to compute than the CBOX lower bound, they must prune more exact EMD calculations than CBOX in order to be as effective in query time.

The queries in the final two examples of this section are both images in the Corel database. The results of the queries



are shown in Figure 5.12 and Figure 5.13, respectively. The distributions for queries C.3.1 and C.3.2 contain 12 and 13 points, respectively. Notice that the brute force query time for the C.3.\* queries is much greater than the brute force query time for the C.1.\* and C.2.\* queries. The difference is that both the query and the database images have a “large” number of points for the C.3.\* queries. All the lower bounds perform well for query C.3.1, but the CBOX lower bound gives the lowest query time. Recall that the CBOX lower bound reduces to the distance between distribution centroids for equal-weight distributions. The centroid distance pruned many exact EMD calculations for C.3.1 because most of the



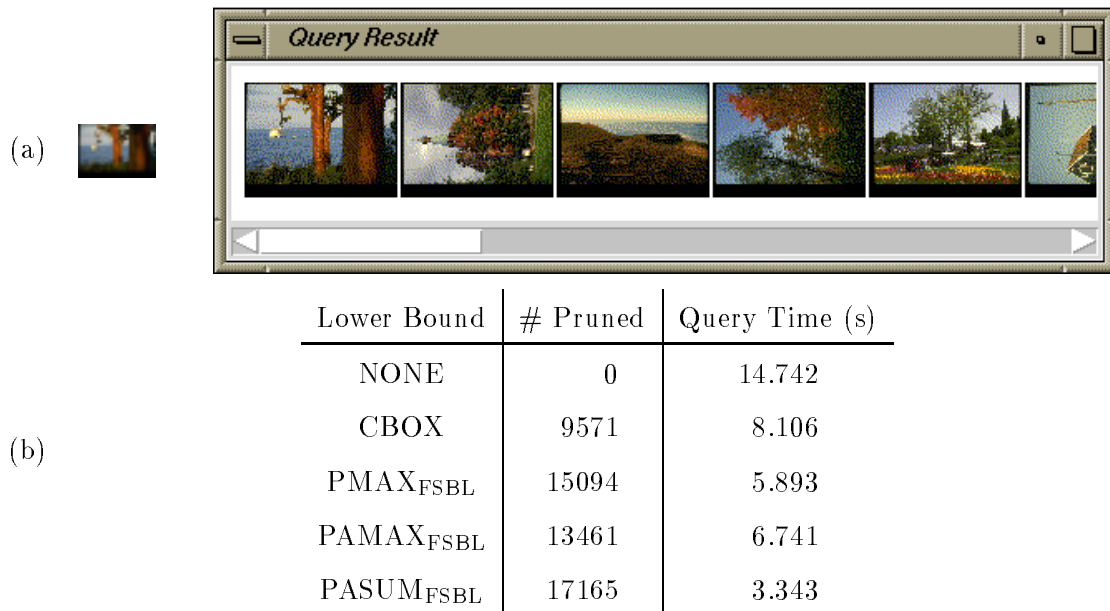


Figure 5.13: Query C.3.2 – Image with Trees, Grass, Water, and Sky. (a) query results. (b) query statistics.

weight in the distribution is around yellow and orange, far from the centroids of the database images (as one can see in Figure 5.11). The blue, green, and brown in query C.3.2 span a larger part of the color space than the colors in C.3.1, the query centroid is close to many database centroids (once again, see Figure 5.11), and the centroid distance lower bound does not perform as well as for C.3.1. The projection-based lower bounds, however, each give a better query time for query C.3.2 than the centroid-distance bound. Recall that the lower bounds  $P_{\text{MAX}}_{\text{FSBL}}$ ,  $P_{\text{AMAX}}_{\text{FSBL}}$ , and  $P_{\text{ASUM}}_{\text{FSBL}}$  reduce to the stronger lower bounds  $P_{\text{MAX}}$ ,  $P_{\text{AMAX}}$ , and  $P_{\text{ASUM}}$  for equal-weight distributions. The  $P_{\text{ASUM}}_{\text{FSBL}}$  lower bound yields a tolerable query time for query C.3.2.

