Web-scale image clustering revisited
Supplementary material

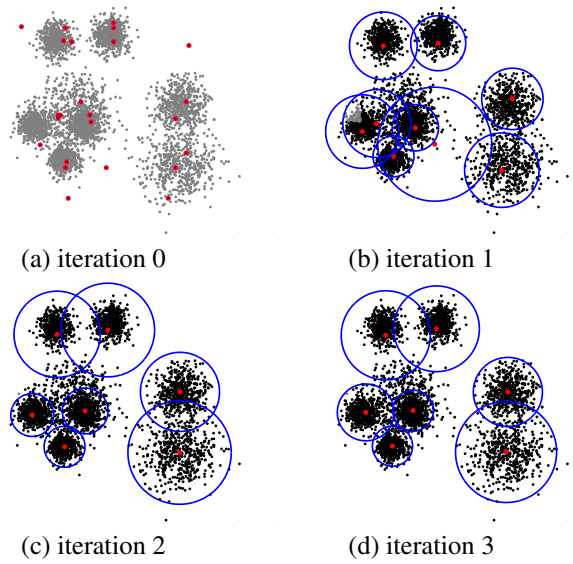
Yannis Avrithis[†], Yannis Kalantidis[‡], Evangelos Anagnostopoulos[†], Ioannis Z. Emiris[†]
[†]University of Athens, [‡]Yahoo! Labs

Abstract

This supplementary material provides additional experiments, discussion and interpretations that complement our paper. The additional results refer to more parameters and options, more datasets, measurements and comparisons, and more large scale experiments.

1. More on dynamic IQ-means

A question arising in the paper is whether the algorithm that purges clusters in dynamic IQ-means is applicable in our case, since EGM assumes a probabilistic Gaussian mixture model, while this work assumes hard assignment as in k-means. It turns out that adequate overlaps can indeed be found indeed under hard assignment. This is illustrated in the small two-dimensional experiment of Fig. 1 that resembles an experiment of EGM, but also shows unassigned points, revealing the search strategy of IQ-means.



(a) iteration 0 (b) iteration 1

(c) iteration 2 (d) iteration 3

Figure 1. First three iterations (0 = initialization) of dynamic IQ-means on a random dataset of $n = 4000$ points following a mixture of 8 normal components, initialized at $k = 24$ with grid size $s = 64$, search window $w = 20$, search target $t_X = 2$, overlap $\tau = 0.6$. Black points: assigned; gray: unassigned; red: centroids; blue circles: 2σ where σ is estimated standard deviation.

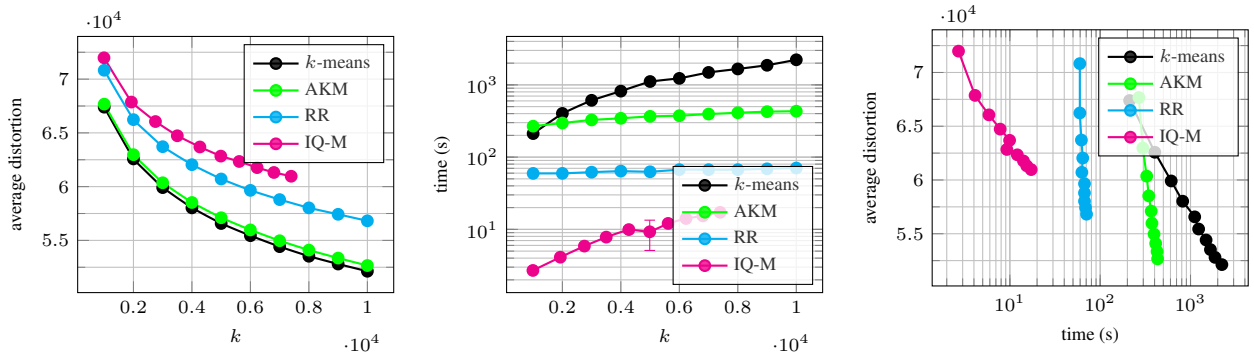
2. More comparisons

Fig. 2 shows average distortion and time measurements just like Fig. 4 in the paper, but for Paris dataset. In particular, Fig. 2a shows that AKM and k-means remain at the same levels of distortion while RR is higher and IQ-means even higher. This discrepancy between RR and IQ-means compared to SIFT1M, along with the fact that a higher number of clusters remain empty for IQ-means, is attributed to the different feature distribution of the two datasets, which would require a finer grid in the case of Paris. It is thus evident that there are two different causes for increased distortion: inverted search from centroids to points (present in both RR and IQ-means) and point quantization (present only in IQ-means).

On the other hand, Fig. 2b is similar to the case of SIFT1M. The ordering of methods remains the same, with IQ-means being the fastest, by at least one order of magnitude for small k. Finally, Fig. 2b is again similar to SIFT1M, but with distortion being slightly higher for IQ-means as discussed above. It is shown more clearly how increased approximation affects both speed (positively) and distortion (negatively). Again, it would be possible to improve IQ-means further using a finer grid.

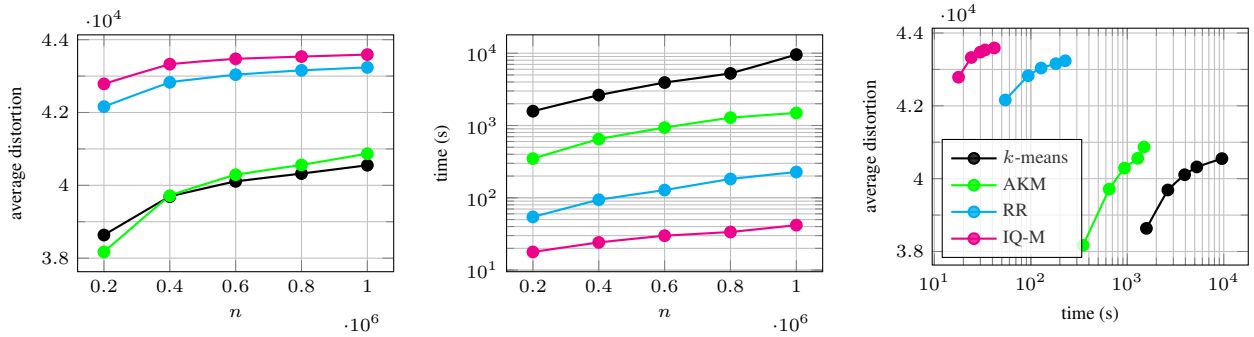
Fig. 3 shows a different experiment for SIFT1M. We now fix k and vary n by clustering increasing subsets of the dataset. Otherwise distortion and time measurements remain the same as in Fig. 2 above and Fig. 4 in the paper. Now both distortion and time are increasing with n for all methods. Again, as shown in Fig. 3a, distortion is similar for k-means and AKM, higher for RR and slightly higher for IQ-means. The latter is only due to some clusters being empty for IQ-means, because otherwise RR and IQ-means exhibit the same behavior in terms of distortion as shown in Fig. 4 in the paper.

Fig. 3b reveals another interesting finding: while all



(a) average distortion *vs*. k (b) time *vs*. k (c) average distortion *vs*. time

Figure 2. Average distortion and total time for 20 iterations on Paris for varying number of clusters k. Time for IQ-means includes encoding of data points that is constant in k, but not codebook learning.



(a) average distortion *vs*. n (b) time *vs*. n (c) average distortion *vs*. time

Figure 3. Average distortion and total time for 20 iterations on SIFT1M for $k = 10^4$ and varying number of data points n. Time for IQ-means includes encoding of data points that is linear in n, but not codebook learning.

methods take more time as n increases, IQ-means remains constant. This is the result of quantizing all points on a fixed grid regardless of n and working on cell distributions alone. The same effect is evident in Fig. 3c, where increased n causes only an increase of distortion for IQ-means, while time remains fixed. The gain in speed varies up to more than two orders of magnitude for k-means.

3. More large scale experiments

Following Section 5.3 in the paper, we report here some further statistics on clustering the YFCC100M and Paris datasets.

Figure 4 is a superset of Table 3 in the paper. We report times per iteration for IQ-means and dynamic IQ-means for different number of clusters and for different values of the overlap parameter τ. The time per iteration for IQ-means grows linearly in the initial number of clusters requested, while clustering in $k = 500$k clusters takes about 11 minutes per iteration. Further speed-up can be achieved by using dynamic IQ-means, where clustering in $k = 500$k centroids takes about 9 minutes per iteration with $\tau = 0.6$ and results in $k' = 356$k final clusters.

As described in Section 5.1 of the paper, we extract noisy

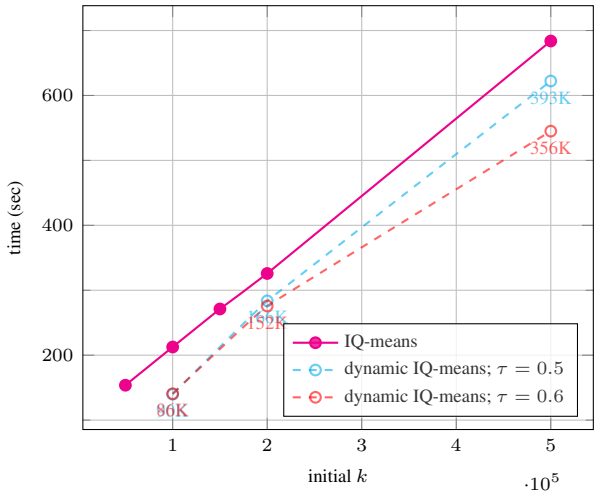


Figure 4. Clustering time per iteration for IQ-means and dynamic IQ-means on the YFCC100M dataset.

labels through automatic image classification. After clustering with IQ-means, for each of the clusters we count how many times each label appears in the images of that cluster. If we keep only the most frequent label for each cluster, we can plot the number of clusters in which a specific label is

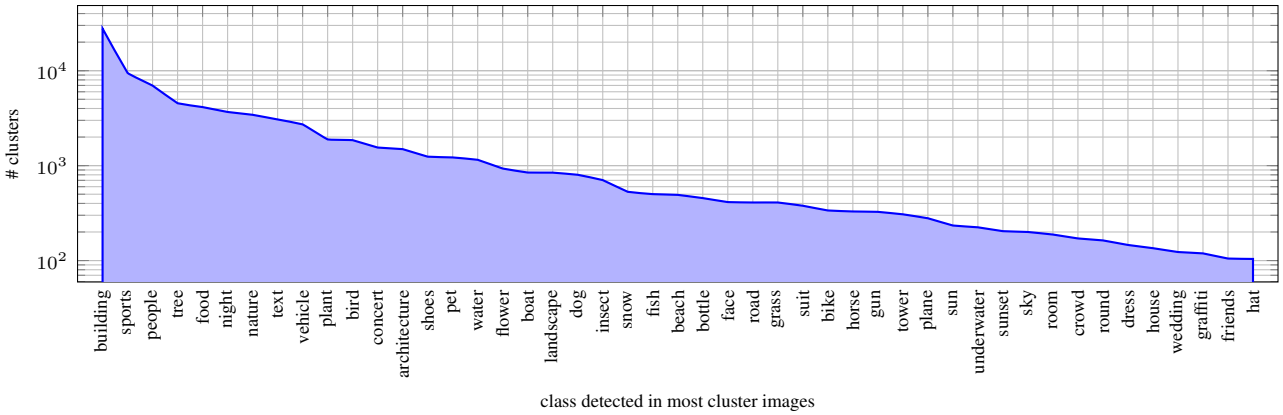


Figure 5. The "most popular class" frequencies. Showing all classes that appear in more than 100 clusters. $k = 10^5$.

the "most popular". In Figure 5 we show exactly this for the top labels, *i.e*. all labels that are the most popular label in at least 100 clusters. As expected, more generic labels like building, sport and people are the most frequent ones, but also specific animals or objects (*e.g*. dog or bottle) appear as most frequent labels in hundreds of clusters each.

Finally, in Figure 6 we show representative images for some of the landmarks of the Paris dataset. We can get such a diverse set without any extra computation, by choosing one image from each one of the clusters the landmark appears in. In particular, we first cluster the dataset with IQ-means and $k = 10^3$. We then select the 12 clusters that contain the most images for each landmark and randomly pick one to display. We see that in most cases there is visible viewpoint, scale or lighting change between images from different clusters, making this sample diverse, despite the approximate nature of our clustering and the global features used.

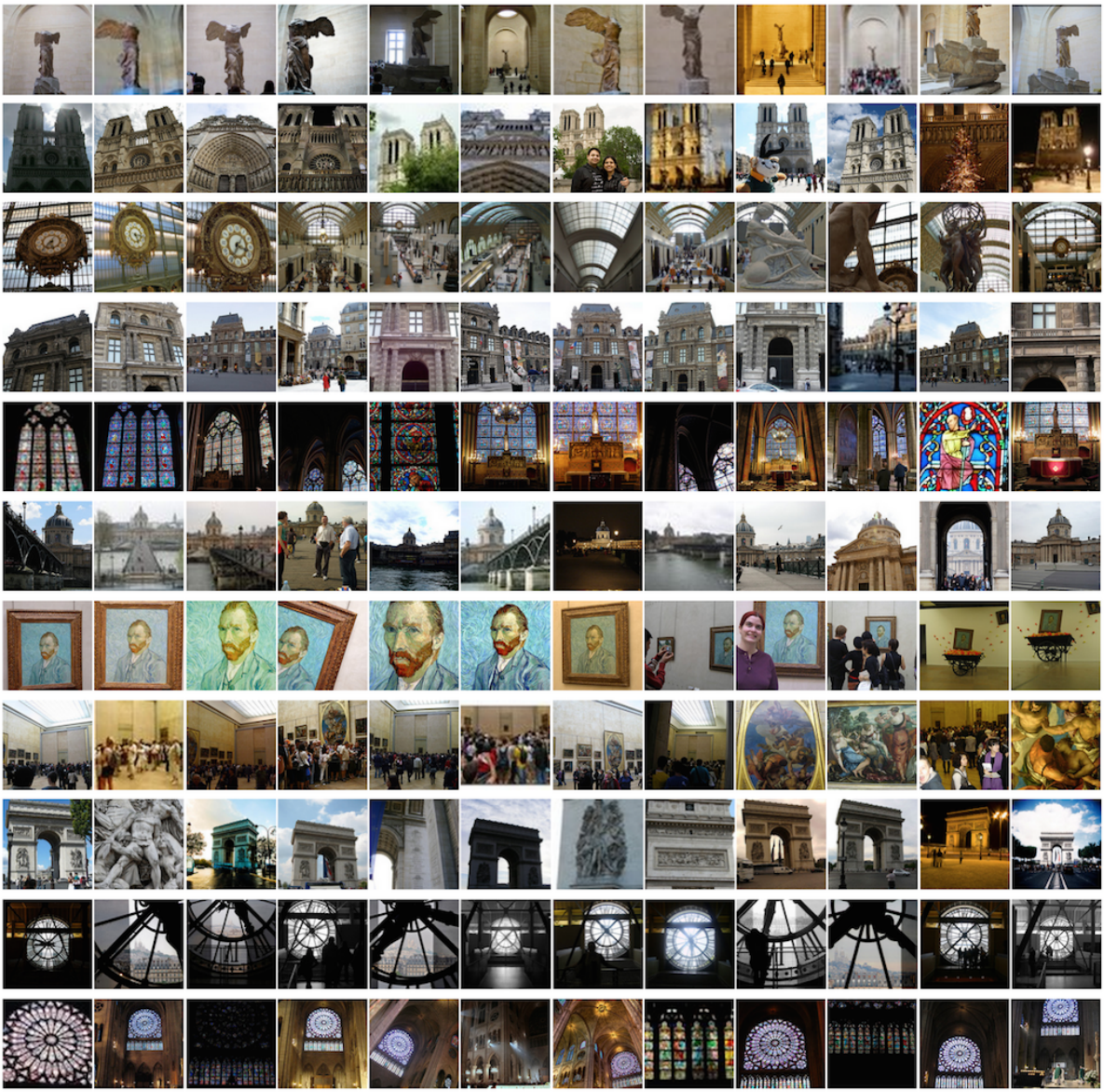


Figure 6. One image from each of the 12 most popular clusters for a number of the Paris dataset landmarks when $k = 10^3$.