"Clustering by Composition" – Unsupervised Discovery of Image Categories Attached Material

Alon Faktor and Michal Irani

Dept. of Computer Science and Applied Math
The Weizmann Institute of Science, ISRAEL

We provide:

1. Proofs for the claims of Section 4.
2. Description of the various Caltech-101 benchmark subsets.
3. The PASCAL-VOC 2010 subset and results.

1. Proofs for the claims of Section 4

Claim 1 [A single shared region between two images]
Let R be a region which is shared by two images, I_1 and I_2 of size N. Then:
(a) Using S random samples per descriptor, guarantees to detect the region R with probability $p \geq \left(1 - e^{-S\,|R|/N}\right)$.
(b) To guarantee the detection of the region R with probability $p \geq (1-\delta)$, requires $S = \frac{N}{|R|}\log(\frac{1}{\delta})$ random samples per descriptor.

Proof: Let R_1 and R_2 denote the instances of a region R in I_1 and I_2. In order to detect the entire region R, at least one descriptor $d_1 \in R_1$ has to randomly sample its correct match $d_2 \in R_2$ (following which the entire region will be 'grown' due to the propagation phase of the Region Growing Algorithm described in Section 4). So, the probability of detecting a region is equal to the probability that at least one of the descriptors $d_1 \in R_1$ will randomly sample its correct match $d_2 \in R_2$.

The probability of a single descriptor $d_1 \in R_1$ to randomly fall on its correct match $d_2 \in R_2$ is $\frac{1}{N}$ (where N is the size of the image). Therefore, the probability that it will NOT fall on d_2 is $(1-\frac{1}{N})$. The probability that NONE of its S samples will fall on d_2 is $(1-\frac{1}{N})^S$. Therefore, the probability that NONE of the descriptors in R_1 will randomly fall on their correct match is $q \triangleq \left(1-\frac{1}{N}\right)^{S|R_1|} = \left(1-\frac{1}{N}\right)^{S|R|}$. Thus the probability of detecting the shared region R is $p \triangleq (1-q)$.

(a) for $N \geq 1$ it holds that $\left(1-\frac{1}{N}\right)^N \leq e^{-1}$. Implying that $q = \left(1-\frac{1}{N}\right)^{N\frac{S|R|}{N}} \leq e^{-\frac{S|R|}{N}}$. So $p = (1-q) \geq 1 - e^{-\frac{S|R|}{N}}$. ■

(b) We need to guarantee that $p = (1-q) \geq 1-\delta$, and ask what is minimal number of samples S required. We know from (a) that $p = (1-q) \geq 1 - e^{-\frac{S|R|}{N}}$. So if we require

$1 - e^{-\frac{S|R|}{N}} \geq 1 - \delta$ we will satisfy the condition. Switching sides we get: $e^{-\frac{S|R|}{N}} \leq \delta$. Applying $\log$ gives us: $-\frac{S|R|}{N} \leq \log(\delta)$. Rearranging the terms: $S \geq \frac{N}{|R|}\log(\frac{1}{\delta})$. ■

Claim 2 [Multiple shared regions between two images]
Let $R_1, \ldots, R_L$ be L shared non overlapping regions between two images I_1 and I_2. If $|R_1| + |R_2| + \ldots |R_L| = |R|$, then it is guaranteed to detect at least one of the regions R_i with the same probability $p \geq (1-\delta)$ and using the same number of random samples per descriptor S as in the case of a single shared region of size $|R|$.

Proof: $R_1, \ldots, R_L$ are non-overlapping regions, so their probabilities of detection are statistically independent of each other. The probability that all of the regions are not detected is therefore equal to the product of the probabilities of each region not being detected: $\prod_{i=1}^{L}\left(1 - \frac{1}{N}\right)^{S|R_i|}$. This is equal to $\left(1 - \frac{1}{N}\right)^{S\sum_{i=1}^{L}|R_i|} = \left(1 - \frac{1}{N}\right)^{S|R|} = q$. So the probability of detecting at least on region is equal to $1 - q$ which is identical to the term obtained in claim 1.a for the probability of detecting a single shared region with size $|R|$. Similarly, we also get the same term for the required number of samples S as was obtained in claim 1.b. ■

We now consider the case of detecting a shared region between a query image and at least one other image in a large collection of M images. For simplicity, let us first examine the case where all the images in the collection are "partially similar" to the query image. We say that two images are **"partial similar"** if they share at least one large region (say, at least 10% of the image size). The shared regions R_i between the query image and each image I_i in the collection may be possibly different ($R_i \neq R_j$).

Claim 3 [Shared regions within an image collection]

Let I_0 be a query image, and let $I_1 \ldots, I_M$ be images of size N which are "partially similar" to I_0. Let $R_1, \ldots, R_M$ be regions of size $|R_i| \geq \alpha N$ such that R_i is shared by I_0 and I_i (the regions R_i may overlap in I_0). Using $S = \frac{1}{\alpha}\log(\frac{1}{\delta})$ samples per descriptor in I_0, distributed randomly across $I_1, .., I_M$, guarantees with probability $p \geq (1 - \delta)$ to detect at least one of the regions R_i.

Proof: We will first develop a term for the probability of not detecting a specific region $R_i (i = 1, \ldots, M)$. The only change from claim 1.a is that the search space is M times larger (since there are M other images instead of only one). So this probability is equal to $(1 - \frac{1}{NM})^{S|R_i|}$. If there were no overlaps between the regions, then the probability $\tilde{q}$ that none of the regions are detected (as was shown in claim 2) equals to the product of the probabilities of each region not being detected:

$\tilde{q} = \prod_{i=1}^{M}(1 - \frac{1}{NM})^{S|R_i|} = \left(1 - \frac{1}{NM}\right)^{S\sum_{i=1}^{M}|R_i|} \leq \left(1 - \frac{1}{NM}\right)^{SM\alpha N} = \left(\left(1 - \frac{1}{NM}\right)^{NM}\right)^{S\alpha} \leq e^{-S\alpha}$.

An overlap between the regions will not change this term. This is due to the fact that on the one hand there are fewer descriptors in the union of all the regions, but on the other hand each descriptor has a higher probability of finding a good match at random. It is easy to show that these two terms cancel each other.

Therefore, the probability of detecting at least one of the regions is equal to $p = (1 - \tilde{q}) \geq \left(1 - e^{-S\alpha}\right)$. Finally, in order to guarantee detection of at least one region with probability $\geq (1 - \delta)$ we need to use $S \geq \frac{1}{\alpha}\log(\frac{1}{\delta})$ samples. ■

Claim 4 [Multiple images vs. Multiple images]
Assume each image in the collection is "partially similar" with at least $\frac{M}{C}$ images (shared regions $\geq 10\%$ image size), and $S = 40C$ samples per descriptor. Then one 'simultaneous iteration' (all the images against each other), guarantees that at least 95% of the images will generate at least one strong connection (find a large shared region) with at least one other image in the collection with high probability. This probability rapidly grows with the number of images, and is practically 100% for $M \geq 500$.

Proof: According to claim 3, in order to guarantee with probability $p \geq 98\%$ that an image I_0 will detect at least one region which is at least 10% of the size of the image and is shared with another image, we are required to use $S = 40$ random samples per descriptors ($\delta = 0.02$ and $\alpha = 0.1$). This is the required number of samples S when all the M images are "partially similar" to I_0. When only $\frac{M}{C}$ of the images are "partially similar" to I_0, then S must be C times larger, i.e. $S = 40C$ (using similar derivations to those in claim 3).

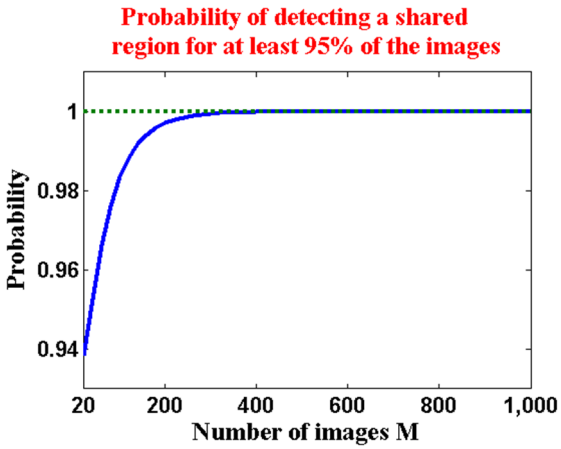
This, however, was for one specific image I_0. When applying this process simultaneously to all the M images, we would like to check what percent of the images will detect with very high probability at least one shared region with another image. We will regard the event of each image trying to detect a shared region as an independent Bernoulli trial with success probability of $p = 0.98$ (the guaranteed probability of detecting a shared region per trial). We have M images, thus M Bernoulli trials, all with the same success probability p. Therefore, The number of successes, i.e., the number of images which detect a shared region, has a Binomial distribution $Bin(M, p)$. Similarly, the number of failures has also a a Binomial distribution $Bin(M, 1 - p)$.

When M is several hundreds ($100 \leq M \leq 1000$) and $1 - p = 0.02$ is quite small, the resulting product $M(1 - p)$ is of an intermediate size (between 2 and 20). It is well known that in these cases, the binomial distribution $Bin(M, 1-p)$ can be approximated well with a Poisson distribution with parameter $\lambda = (1 - p)M = 0.02M$. In other words, the probability that k images did not detect a shared region can be approximated by $\frac{e^{-\lambda}(\lambda)^k}{k!}$.

The probability that at least rM of the images detected at least one shared region is equal to the probability that all the images detected a region ($k = 0$), or that all but one

detected a shared region ($k = 1$), ... or that all but $(1 - r)M$ detected a shared region. Therefore, it can be approximated by $\sum_{k=0}^{(1-r)M} \frac{e^{-\lambda}(\lambda)^k}{k!}$.

The graph below shows this probability for $r = 95\%$ as function of the number of images M. We can see that the probability that at least 95% of the images detected a shared region is very high and goes to 1 as M increases (is practically 100% for $M \geq 500$). ■



2. Description of the various Caltech-101 benchmark subsets

We used the following benchmark subsets of Caltech-101 (this is the experimental setting defined by [7]):

1. 7 classes - Faces (first 100 images), Motorbikes (first 100 images), Dollar bill (52), Garfield (34), Snoopy (35), Stop sign (64) and Windsor chair (56). These are a **total of 441 images**.
2. 20 classes - Faces (first 100 images), Motorbikes (first 100 images), Dollar bill (52), Garfield (34), Snoopy (35), Stop sign (64), Windsor chair (56), Leopards (first 100 images), Binocular (33), Brain (98), camera (50), Car side (first 100 images), ferry (67), Hedgehog (54), Pagoda (47), Rhino (59), Stapler (45), Water Lilly (37), Wrench (39) and Yin yang(60). These are a **total of 1230 images**.
3. 4 classes - Faces (first 50 images), Dalmatians (first 50 images), Hedgehogs (first 50 images) and Okapi (39). These are a **total of 189 images**.
4. 10 classes - Faces (first 50 images) ,Dalmatians (first 50 images), Hedgehogs (first 50 images), Okapi (39), Leopards (first 50 images), Car Side (first 50 images), Cougar Face (first 50 images), Guitar (first 50 images), Sunflower (first 50 images) and Wheelchair (first 50 images). These are a **total of 489 images**.

Bellow are some typical example images from these subsets:

3. The PASCAL-VOC 2010 subset and results

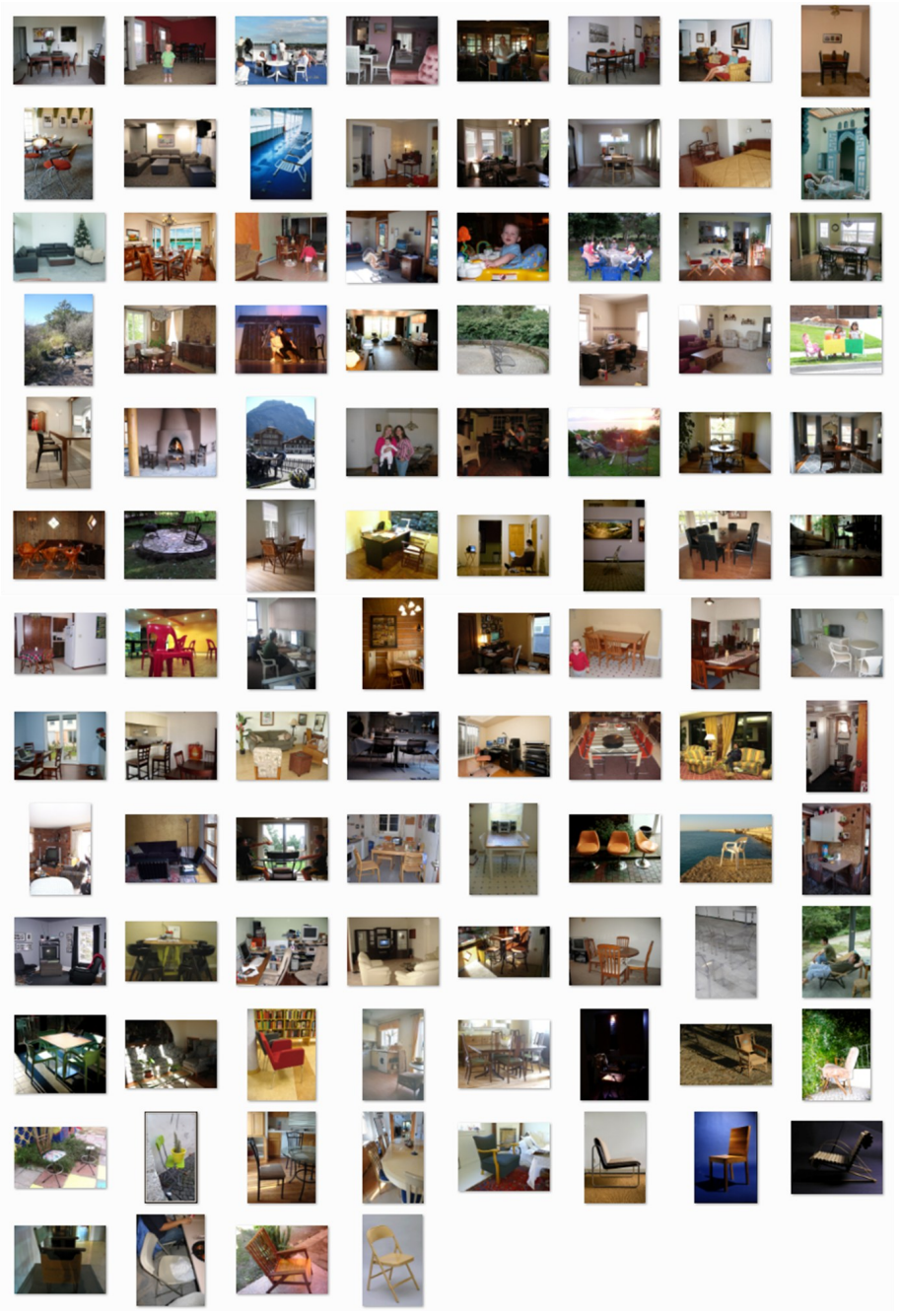
We next present the PASCAL-VOC 2010 subset used in our experiments, followed by our clustering results on the subset. The subset contains 100 images per category (cars, bicycles, horses and chairs). The object size in each category ranges from 5% of the image size to almost the entire image. The full subset is provided next:

Cars (100 images):

Bicycles (100 images):

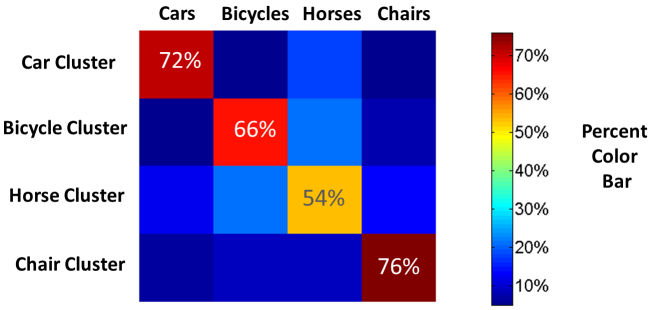
Horses (100 images):

Chairs (100 images):



Clustering results

We ran our clustering algorithm on this subset and obtained the following confusion matrix:



Each row in the confusion matrix represents one cluster and the columns show the distribution of images from different categories within the cluster. Ideally, we would like the values on the diagonal to be 100% and the off-diagonal values to be 0%.

As can be seen, the strongest confusion is between Horses and Bicycles. The reason for this confusion is due to the very unique yet almost identical pose of the human riding these two types of objects. Some examples of this can be seen in the figure below:

Such a similar non-trivial pose of the rider induces strong affinities between those images, thus resulting in confusion between those two categories.

Similarly, there is some confusion between Cars and Horses. This is because several of the horse images contain carriages, which resemble cars, inducing strong affinities to car images. Some examples of this can be seen in the figure below:

Nevertheless, observing the graph below, we can see that most of the mis-clustered images (marked by a red X on the graph) have equally good affinities to two different clusters (the second cluster usually being their correct cluster). The N-cut algorithm forces them into one of the clusters - unfortunately, the wrong one. In contrast, most of the well-clustered images (marked by a green O on the graph) have strong affinity to only one cluster – their correct cluster (and weak affinity to all other clusters). This indicates that replacing N-cuts with an iterative refinement of the final clustering assignment can significantly improve our clustering results. This is part of our future work.