Texture-Based Image Retrieval Without Segmentation

Yossi Rubner and Carlo Tomasi
Computer Science Department
Stanford University
Stanford, CA 94305
[rubner,tomasi]@@cs.stanford.edu

Abstract

Image segmentation is not only hard and unnecessary for texture-based image retrieval, but can even be harmful. Images of either individual or multiple textures are best described by distributions of spatial frequency descriptors, rather than single descriptor vectors over presegmented regions. A retrieval method based on the Earth Movers Distance with an appropriate ground distance is shown to handle both complete and partial multi-textured queries. As an illustration, different images of the same type of animal are easily retrieved together. At the same time, animals with subtly different coats, like cheetahs and leopards, are properly distinguished.

1. Introduction

Perceptually adequate descriptors of image texture are important cues for image retrieval [14, 15, 8, 13, 2, 21] when used in combination with other descriptors like color and shape. Gabor filters [9] where shown to be perceptually plausible texture descriptors [6], and, for individual textures, the Earth Movers distance (EMD) [19, 17] was shown to both match perceptual similarity well and tolerate variations in orientation, scale, illumination, and other sources of changes in texture appearance.

In image retrieval, however, multi-texture queries are to be compared to multi-texture images. How can individual texture descriptors and distances for texture comparison be lifted, so to speak, to the level of distances between entire *distributions* of textures descriptors? To make this task even harder, queries usually specify *partial* image content: one looks, say, for a cheetah chasing a zebra (multiple textures) without regard for the background (partial query). The standard answer [14, 8, 13, 11, 3] to these questions relies on texture segmentation. The images are first split into regions of uniform textures, and a similarity measure that compares individual textures is then applied between pairs of such regions. Thus, the similarity of two images will be determined by some combination of the similarities between pairs of regions in the two images. Two major problems with this approach are the following:

- Texture segmentation is hard, and the notion of "uniform texture" that it implies is not well defined. In addition, different segmentations may be plausible at different scales of resolution. For example, each leaf in a tree might be segmented out at one scale, while the whole tree-top or the whole forest might be considered to be individual regions at coarser scales. A mismatch between query and image descriptors in terms of resolution scales may lead to retrieval errors.

- Retrieval based on comparing texture segments is usually sensitive to over- and under-segmentation. On one hand, spatial changes in texture appearance can cause single textures to be split into smaller segments (over-segmentation). On the other hand, the segmentation algorithm can mistakenly combine together small regions of different textures (under-segmentation). In addition, problems may occur when some texture in the image, although significant in size when combined together, is scattered over the image and therefore lost. An example of this phenomenon is an aerial view of a town in a richly vegetated area, in which both buildings and vegetation are made up of numerous but small texture patches.

In this paper, we show that the EMD, which worked well for individual texture descriptors, can be used once more, at a higher level, to address all these difficulties at once, and that partial and multi-texture queries can be answered well on a database of complex, natural images without performing any segmentation. This is possible because of the EMD's built-in ability to find appropriate correspondences in texture space between the elements that comprise the query's texture distribution on one hand, and the image's distribution on the other.

2. Distributions in Texture Space

Texture involves a strong notion of spatial extent: a single point has no texture. If texture is defined in the frequency domain, the texture information of a point in the image is carried by the frequency content of a neighborhood of it. Gabor functions are commonly used in texture analysis to capture this information (e.g. [4, 7, 13]). There is also evidence that simple cells in the primary visual cortex can be modeled by Gabor functions tuned to detect different orientations and scales on a log-polar grid [6].

In this paper we used a similar dictionary of Gabor filters as the one derived in [13][1]. Applying these Gabor filters to an image results for every image location in a texture vector

$$\mathbf{t} = [t_1, \ldots, t_d]^T \ , \qquad (1)$$

where d is the number of scales times the number of orientations that are used in the filter dictionary. We used four scales and six orientations so that $d = 24$.

These *texture features* reflect the components of the texture in terms of scales and orientations. Figure 1 shows an example of a texture feature. Part (b) shows the spatial average of each of the 24 filter responses over the image in part (a) of the figure. Darker squares represent stronger responses. Notice the two strong responses that correspond to the vertical and horizontal texture components at an intermediate scale.

The texture content of an entire image is represented by a distribution of texture features, a cloud of points in a space of 24 dimensions. This distribution accounts for four sources of variation in the filter responses:

1. The size of the basic texture element ("texton") is often larger than the support of at least the finest scale Gabor filters. This causes a variation in the filter responses even within textures that a human would perceive as homogeneous in the image. To address this variation, many texture analysis methods (see for instance [1, 22, 12]) integrate filter responses over areas that are larger than the largest filter support.

2. Texture regions that a human would perceive as being homogeneous *in the world* can produce inhomogeneous regions *in the image* because of foreshortening and variations in illumination. This spreads the distribution in texture space and increases its variability.

3. Textures exhibit spatial variation even in the world. For instance, most natural textures are regular only in a statistical sense, so filter responses will vary regardless of viewing conditions.

4. Images with multiple textures result in a combination of the distributions of the constituent textures.

Because of all these sources of variation, a single image can produce nearly as many 24-dimensional texture vectors as it has pixels. To represent the full distribution of image texture in a compact way, we first find the dominant clusters in the 24 dimensional texture space by using a similar clustering algorithm as the one used for color in [18]. This algorithm returns a variable number of clusters depending of the complexity of the distribution. While this method is simple and fast, so that large number of images can be processed quickly, more sophisticated clustering algorithms (e.g., see [3]) may further improve our texture similarity methods. The resulting set of cluster centers together with the fractional cluster weights is the *texture signature* of the image. An example of a texture signature with four clusters is shown in Figure 1(c).

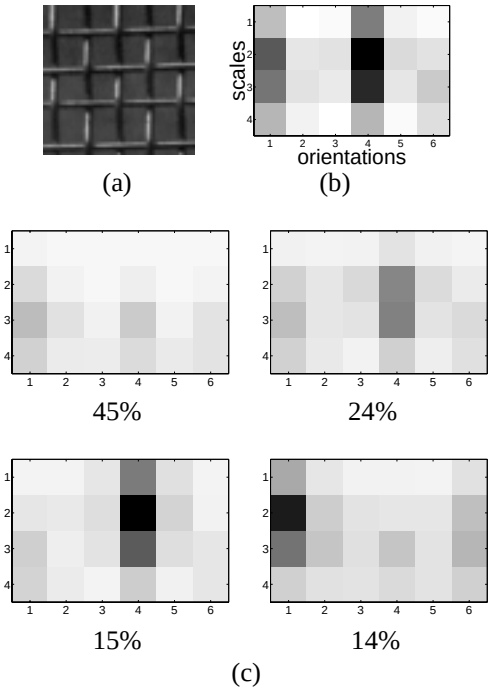


Figure 1. (a) Texture patch from [5]. (b) Average over all texture features. The Gabor filter bank consists of four scales and six orientations. (c) The four clusters in the texture signature together with their weights (in percentage of the number of pixels).

3. Texture Distance

In [19] the *Earth Mover's Distance* (EMD) is introduced as a flexible similarity measure between multidimensional distributions, and is described in detail in [17]. Intuitively, given two distributions represented by signatures, one can be seen as a mass of earth properly spread in space, the other as a collection of holes in that same space. Then, the EMD

[1] The full derivation of our Gabor filters can be found in [17]

measures the least amount of work needed to fill the holes with earth. Here, a unit of work corresponds to transporting a unit of earth by a unit of *ground distance*, which is a distance in the feature space. The EMD is based on the transportation problem [10] and can be solved efficiently by linear optimization algorithms that take advantage of its special structure.

Formally, let $P = \{(\mathbf{p}_1, w_{\mathbf{p}_1}), \ldots, (\mathbf{p}_m, w_{\mathbf{p}_m})\}$ be the first signature with m clusters, where $\mathbf{p}_i$ is the cluster representative and $w_{\mathbf{p}_i}$ is the weight of the cluster; $Q = \{(\mathbf{q}_1, w_{\mathbf{q}_1}), \ldots, (\mathbf{q}_n, w_{\mathbf{q}_n})\}$ the second signature with n clusters; and $\mathbf{DIST} = [\text{dist}(\mathbf{p}_i, \mathbf{q}_j)]$ the ground distance matrix where $\text{dist}(\mathbf{p}_i, \mathbf{q}_j)$ is the distance between clusters $\mathbf{p}_i$ and $\mathbf{q}_j$. The EMD between signatures P and Q is then

$$\text{EMD}(P, Q) = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}\,\text{dist}(\mathbf{p}_i, \mathbf{q}_j)}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \ . \qquad (2)$$

where $\mathbf{F} = [f_{ij}]$, with $f_{ij} \geq 0$ the flow between $\mathbf{p}_i$ and $\mathbf{q}_j$, is the optimal admissible flow from P to Q that minimizes the numerator of (2) subject to the following constraints:

$$\sum_{j=1}^{n} f_{ij} \leq w_{\mathbf{p}_i} \ , \qquad \sum_{i=1}^{m} f_{ij} \leq w_{\mathbf{q}_j}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min\Big(\sum_{i=1}^{m} w_{\mathbf{p}_i}, \sum_{j=1}^{n} w_{\mathbf{q}_j}\Big) \ .$$

In this work we take advantage of the following properties of the EMD:

• Adaptive representation of high-dimensional distributions of features for each image independently (in contrast to other methods that either use one global adaptive representation based on the combined distributions of all the images together, or represent only the one-dimensional marginals of the full distribution [13, 16]).

• Robustness to small variations of feature values.

• No need for explicit segmentation. The representation by a finite number of clusters does not suffer from over- and under-clustering. Also, splitting a cluster into few sub-clusters will not significantly change the EMD results as long as the sub-clusters are mutually close in the feature space.

• Partial matches can be done in a very natural way. This is important, for instance, for image retrieval and in order to deal with occlusions and clutter, and is illustrated in section 4 below.

More details on the EMD can be found in [17].

For our texture signatures, we use the following ground distance:

$$\text{dist}(\mathbf{p}, \mathbf{q}) = 1 - e^{-\frac{\|\mathbf{p}-\mathbf{q}\|_1}{D}} \ , \qquad (3)$$

where $\mathbf{p}$ and $\mathbf{q}$ are two texture vectors as in (1), $\|\cdot\|_1$ is the L_1 norm, and D is a constant that distinguishes between "close" and "far" distances in the feature space. In this paper we use

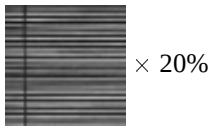
$$D = d(\mathbf{0}, \frac{1}{2}\boldsymbol{\sigma}) \ ,$$

where $\mathbf{0}$ is the zero vector, $\boldsymbol{\sigma} = [\sigma_1 \ldots \sigma_d]^T$ is a vector of standard deviations of the components of the features in each dimension from the overall distribution of all images in the database, and d is the dimensionality of the feature space. Assuming that the distribution of the features is unimodal, D is a measure of the spread of the distribution. The bigger the spread, the larger the distances are, in general. This saturated ground distance agrees with results from psychophysics. In [20] it is argued that the similarity between stimuli of any type can be expressed as *generalization data* by $g(\delta(S_i, S_j))$, where δ is a perceptual distance between two stimuli, the L_1 norm in our case, and g is a generalization function such as $g(\delta) = \exp(-\delta^\tau)$. This is equivalent to our dissimilarity measure which can be expressed in term of the similarity $g(\delta)$ by $1 - g(\delta)$. The rationale for this distance measure is that only texture descriptors that closely agree in several of their components are deemed to be close to each other. In a space with 24 dimensions, any metric that increases with point separation as fast as the L_2 norm (Euclidean distance) or even as fast as the L_1 norm (Manhattan distance) is bound to give poor retrieval results. This is because with such metrics distances between unrelated features are very large. As a consequence, if a distribution of descriptors from a single texture is marred by a few outliers, the EMD between similar textures will be dominated by a small number of large flow values. With our definition, ground distances are saturated to 1, and numerous agreements between individual elements in two similar distributions can still cause the EMD to be relatively small. In addition, since the constant D in (3) is based on a statistical parameter of the database at hand, our ground distance adapts naturally to the range of variation within the database itself.

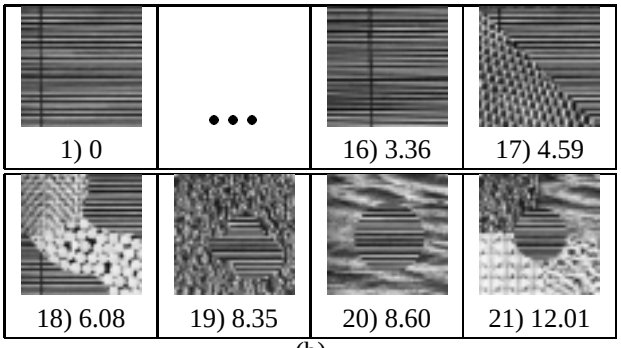
4. Partial Matches

Having defined texture signatures and a ground distance between them, we can now use the EMD to retrieve images with textures. Here we demonstrate the ability to handle images that contain more than one texture without first segmenting the images.

In the first experiment we constructed a database of 1792 texture patches, by dividing each of 112 textures from the Brodatz album [5], into 4 by 4 non-overlapping 128 by 128 pixel patches. To this database we added images that were composed by mosaicing together several texture patches from the database. After the clustering process, the average size of the texture signatures was 12 clusters per patch.



(a)



(b)

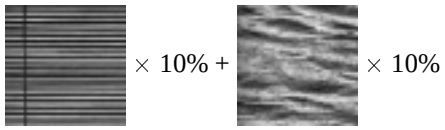
Figure 2. Partial texture query: 20% of the texture in part (a) and 80% "don't care." (b) The 21 best matches: 16 patches from the same texture (only the first and last are shown), followed by all the compositions that contain some part of the queried texture.

Figure 2 shows an example of a partial query. The query was 20% of the texture in part (a) and 80% "don't care." The best matches are shown in part (b) with the 16 patches from the same texture at the beginning followed by all the compositions with some part of the queried texture. We emphasize that no segmentation was performed. Since the total weight of the signature is only 20%, the EMD will return as good matches also images with relatively small amount of the queried texture. Figure 3 demonstrates a partial query where the query has more than one texture.

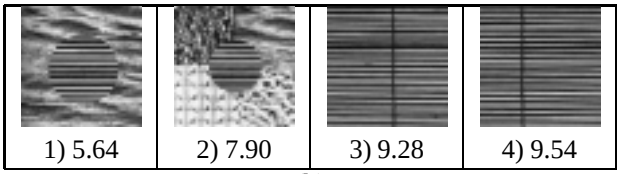
5. Retrieving Natural Images

In the next experiment we created a database of 500 grayscale images of animals from the Corel Stock Photo Library[2] with image sizes of 768-by-512 pixels. We preprocessed the images by our clustering procedure, and obtained an average signature size of 32 clusters. Since most of the queries consists of a single, or a few textures, their signatures are significantly smaller and the EMD computation is more efficient.

[2]The Corel Stock Photo Library consists of 20,000 images organized into sets of 100 images each. We used the following sets: 123000 (Backyard Wildlife), 134000 (Cheetahs, Leopards & Jaguars), 130000 (African Specialty Animals), 173000 (Alaskan Wildlife), and 66000 (Barnyard Animals).



(a)



(b)

Figure 3. Another partial query. The query now contains 10% of each of the two patches in part (a) and 80% "don't care." (b) The two best matches are the two compositions that contain the textures in the query, followed by the patches that contain only one of the queried textures.

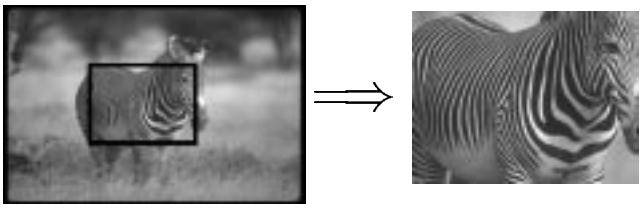
Figure 4(a) shows an example of a query that used a rectangular patch from an image of a zebra. We asked for images with at least 20% of this texture. The 16 best matches (Figure 4(b) shows the 12 most similar to the query) are all images of zebras, out of a total of 34 images of zebras in the database. The various backgrounds in the retrieved images were ignored by the system because of the EMD's ability to handle partial queries. Notice also that in some of the retrieved images there are a few small zebras, which provide a significant amount of "zebra texture" only when combined together. Methods based on segmentation are likely to have problems with such images.

Next we searched for images of cheetahs. The database has 33 images of cheetahs, and 64 more images of leopards and jaguars that have similar texture as cheetahs. Figure 5 shows the query and the 12 best matches. The first eight images are indeed cheetahs. The other four matches are images of leopards and jaguars.

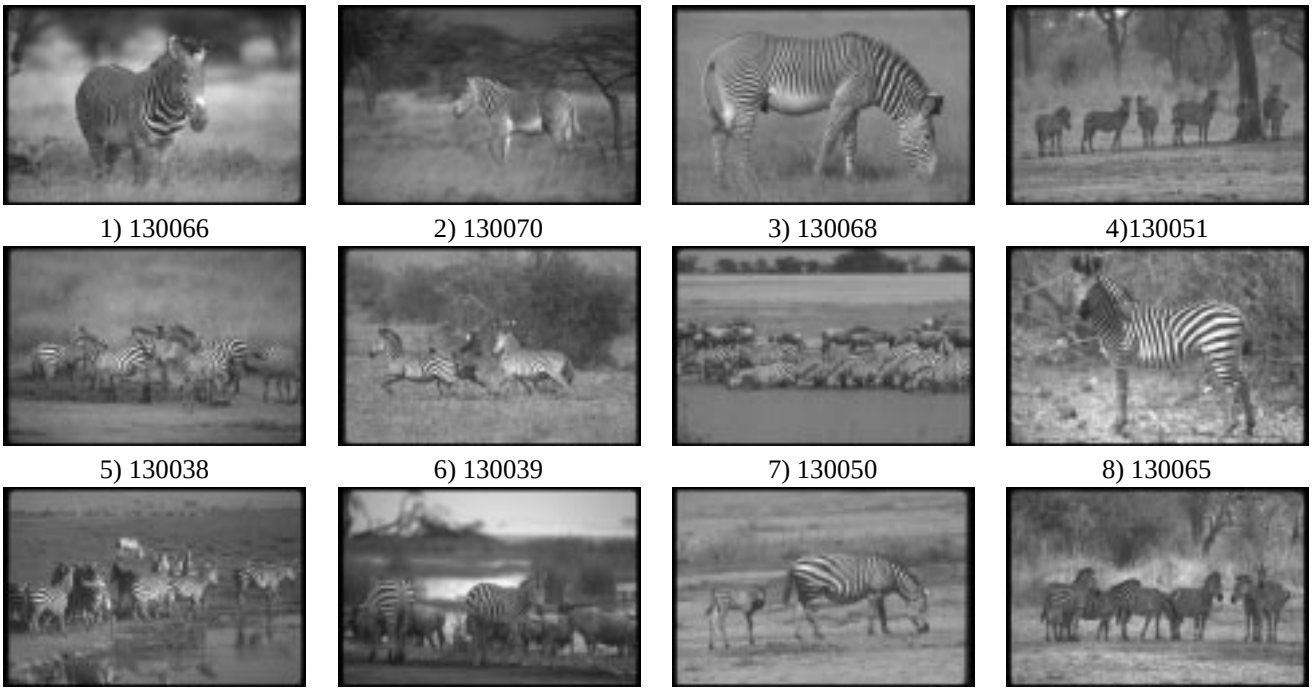
To check if our method can distinguish between the different families of wild cats, we looked for images of jaguars. Figure 6 shows the query results. From the best twelve matches, eleven are jaguars and leopards, which are almost indistinguishable. Only the sixth match was an image of a cheetah.

6. Conclusions

The main point of this paper is that segmentation is not only unnecessary for texture-based image retrieval, but can even be harmful. Because of variations of appearance both within the same texture and across different textures in an



(a)



1) 130066 2) 130070 3) 130068 4)130051

5) 130038 6) 130039 7) 130050 8) 130065

9) 130054 10) 130043 11) 130055 12) 130046

(b)

Figure 4. Looking for zebras. (a) An image of a zebra and a block of zebra stripes extracted from it. (b) The best matches to a query asking for images with at least 10% of the texture in (a). The numbers in the thumbnail captions are indices into Corel CDs.

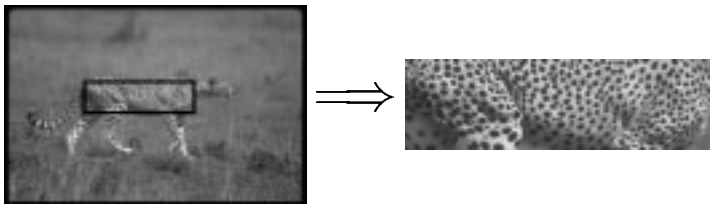
image, both individual textures and multi-textured images are best described by distributions of descriptors, rather than by individual descriptors. We have proposed an effective implementation of this principle by replacing image segmentation with clustering of similar texture descriptors into compact, but detailed and versatile texture *signatures*. The Earth Movers Distance, together with an appropriate ground distance, has proven effective in handling queries, both complete and partial, in a small but difficult test database. Retrieval based on these ideas can handle the wide variations between very different images of related subjects, and can for instance retrieve images of zebras regardless of their number, sizes, backgrounds, or viewing conditions. At the same time, the proposed retrieval techniques can distinguish between rather subtly different images, and can for instance tell cheetahs apart from leopards and jaguars.

Of course, more quantitative experiments with larger databases are in order. However, we believe that we have provided a new and effective approach to a difficult problem in image retrieval.

References

[1] E. H. Adelson and J. R. Bergen. Spatiotemporal energy models for the perception of motion. *JOSA A* , 2(2):284–299, 1985.

[2] J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, and C. Shu. Virage image search engine: an open framework for image management. In



(a)

1) 134000 2) 134008 3) 134012 4) 134024

5) 134025 6) 134027 7) 134028 8) 134029

9) 134040 10) 134052 11) 134057 12) 134073

(b)

Figure 5. Looking for cheetahs. (a) The query. (b) The best matches with at least 10% of the query texture. The last four images are leopards and jaguars which have similar texture as cheetahs. However, cheetahs come first.

SPIE Conf. on Storage and Retrieval for Image and Video Databases IV, 2670: 76–87, 1996.

[3] S. Belongie, C. Carson, H. Greenspan, and J. Malik. Color- and texture-based image segmentation using EM and its application to content-based image retrieval. In *ICCV*, pages 675–682, 1998.

[4] A. C. Bovik, M. Clark, and W. S. Geisler. Multichannel texture analysis using localized spatial filters. *IEEE PAMI*, 12(12):55–73, 1990.

[5] P. Brodatz. *Textures: A Photographic Album for Artists and Designers*. Dover, New York, NY, 1966.

[6] J. D. Daugman. Complete discrete 2-d Gabor transforms by neural networks for image analysis and compression. *IEEE ASSP*, 36:1169–1179, 1988.

[7] F. Farrokhnia and A. K. Jain. A multi-channel filtering approach to texture segmentation. In *CVPR*, pages 364–370, 1991.

[8] D. Forsyth, J. Malik, M. Fleck, H. Greenspan, and T. Leung. Finding pictures of objects in large collections of images. In *Int'l Workshop on Object Recognition for Computer Vision*, Cambridge, UK, 1996.

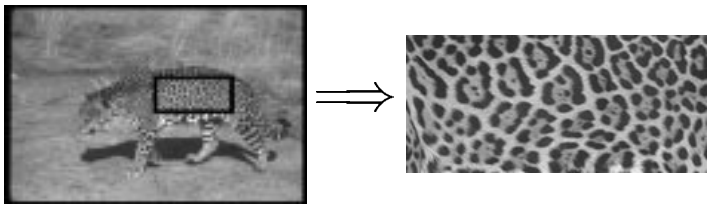
[9] D. Gabor. Theory of communication. *The Journal of the IEE, Part III*, 93(21):429–457, 1946.

[10] F. L. Hitchcock. The distribution of a product from several sources to numerous localities. *J. Math. Phys.*, 20:224–230, 1941.

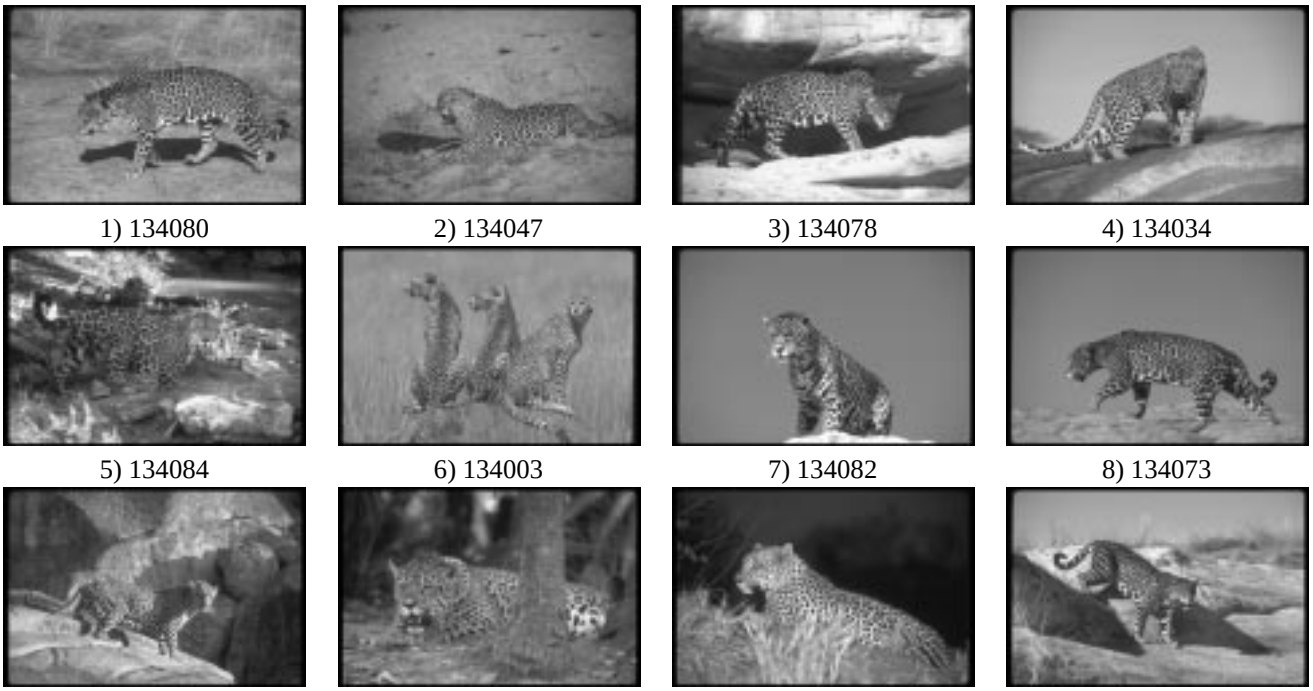
[11] W. Y. Ma. *NETRA: A Toolbox for Navigating Large Image Databases*. PhD thesis, University of California at Santa Barbara, 1997.

[12] J. Malik and P. Perona. Preattentive texture discrimination with early vision mechanisms. *JOSA A*, 7(5):923–932, 1990.

[13] B. S. Manjunath and W. Y. Ma. Texture features for browsing and retrieval of image data. *IEEE PAMI*, 18(8):837–842, 1996.



(a)



1) 134080 2) 134047 3) 134078 4) 134034

5) 134084 6) 134003 7) 134082 8) 134073

9) 134044 10) 134090 11) 134050 12) 134070

(b)

Figure 6. Looking for leopards and jaguars. (a) The query. (b) The best matches with at least 10% of the query texture. All but the sixth image are leopards and jaguars. The sixth image is of cheetahs.

[14] W. Niblack, R. Barber, W. Equitz, M. D. Flickner, E. H. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, G. Taubin, and Y. Heights. Querying images by content, using color, texture, and shape. In *SPIE Conf. on Storage and Retrieval for Image and Video Databases*, 1908:173–187, 1993.

[15] A. Pentland, R. W. Picard, and S. Sclaroff. Photobook: content-based manipulation of image databases. *IJCV*, 18(3):233–254, 1996.

[16] J. Puzicha, T. Hofmann, and J. Buhmann. Non-parametric similarity measures for unsupervised texture segmentation and image retrieval. In *CVPR*, pages 267–272, 1997.

[17] Y. Rubner. *Perceptual Metrics for Image Database Navigation*. PhD thesis, Stanford University, 1999.

[18] Y. Rubner, L. J. Guibas, and C. Tomasi. The earth mover's distance, multidimensional scaling, and color-based image retrieval. In *ARPA IU Workshop*, pages 661–668, 1997.

[19] Y. Rubner, C. Tomasi, and L. J. Guibas. A metric for distributions with applications to image databases. In *ICCV*, pages 59–66, 1998.

[20] R. N. Shepard. Toward a universal law of generalization for psychological science. *Science*, 237:1317–1323, 1987.

[21] J. R. Smith. *Integrated Spatial and Feature Image Systems: Retrieval, Analysis and Compression*. PhD thesis, Columbia University, 1997.

[22] H. Voorhees and T. Poggio. Detecting textons and texture boundaries in natural images. In *ICCV*, pages 250–258, 1987.