MPEG-7: THE MULTIMEDIA CONTENT DESCRIPTION INTERFACE

Peter Stanchev, David Green Jr., and Boyan Dimitrov

Abstract: In this paper a review of the most used MPEG-7 descriptors are presented. Some considerations for choosing the most proper descriptor for a particular image or video data set are outlined.

Keywords: MPEG-7, Multimedia, Content based retrieval

1. Introduction

More and more digital images and video are being captured and stored. In order to use this information, an efficient retrieval technique is required. One major development in this area is the content based image and video retrieval techniques which use image features for indexing and retrieval [Rabitti, 1989]. The most used features are color, texture, and shape. Several semantic image and video models are suggested [Stanchev, 1999], [Grosky, 2001]. In MPEG-7 standard different descriptors for this purpose are proposed [Manjunath, 2002]. What descriptor is the best for a particular data set? Some preferable answers of this question are given.

2. MPEG-7 Descriptors

The MPEG-7 descriptors can be classified as general visual descriptors and domain specific descriptors. The former include color, texture, shape and motion features. The latter includes face recognition descriptor. Although distance functions are not part of the standard, we will present the most used distance functions. Only color, texture and shape descriptors are covered, since they are used mostly.

2.1. Color descriptors

Color is one of the most widely used image and video retrieval features [Schettini, 2001]. The MPEG-7 standard includes five color descriptors which represents different aspects of the color and includes color distribution, spatial layout, and spatial structure of the color. The histogram descriptors capture the global distribution of colors. The dominant color descriptor represents the dominant colors used. The color layout descriptor captures the spatial distribution or layout of the colors in a compact representation. While MPEG-7 standards accommodate different color spaces, most of the color descriptors are constrained to one or a limited number of color spaces for ensuring inter-operability.

2.1.1. Dominant color descriptor

This descriptor specifies a set of dominant colors in an image [Cieplinski, 2000]. It is good to represent color features where a small number of colors are enough to characterize the color information. The extraction algorithm quantizes the pixel color values into a set of dominant colors. The matching is done by calculating the distances between dominant color sets based on the difference between corresponding colors in any two sets of dominants.

The result of the method is a vector with integer numbers, presented as $F = \{(c_i, p_i, v_i), s\}$, (*i=1,2, …, N*), where *N* is the number of dominant colors. The vector components are: the dominant color value c_i (RGB color space vector); p_i - normalized fraction of pixels corresponding to color c_i; optimal color variance v_i, (describes the variance of the color values of the pixels in a cluster around the corresponding color); and the coherency s representing the overall spatial homogeneity of the dominant colors.

The distance algorithm uses an estimate of the mean square error, based on the assumption that the sub-distributions described by dominant colors and variances are Gaussian. Consider 2 descriptors:

$$F_1 = \left\{\left(c_1^{'}, p_1^{'}, v_1^{'}\right), s_1^{'}\right\} \qquad (i = 1,2,..., N_1) \quad \text{and}$$

$$F_2 = \left\{\left(c_2^{'}, p_2^{'}, v_2^{'}\right), s_2^{'}\right\} \qquad (i = 1,2,..., N_2)$$

where $p \in [0,31]$, $c_i = rgb2luv(c_i^{'})$, $v_i = \begin{cases} 60.0 & v_i^{'} = 0 \\ 90.0 & v_i^{'} = 1 \end{cases}$, $p_i = \frac{(p_i^{'} + 0.5)/31.9999}{\sum_i p_i}$, and if

$$f_{x_i y_j} = \frac{1}{2\pi\sqrt{2\pi \times (v_{x_i}^{(l)} + v_{y_j}^{(l)}) \times (v_{x_i}^{(u)} + v_{y_j}^{(u)}) \times (v_{x_i}^{(v)} + v_{y_j}^{(v)})}} \times \exp\left\{-\frac{1}{2}\left[\frac{(c_{x_i}^{(l)} - c_{y_j}^{(l)})^2}{v_{x_i}^{(l)} + v_{y_j}^{(l)}} + \frac{(c_{x_i}^{(u)} - c_{y_j}^{(u)})^2}{v_{x_i}^{(u)} + v_{y_j}^{(u)}} + \frac{(c_{x_i}^{(v)} - c_{y_j}^{(v)})^2}{v_{x_i}^{(v)} + v_{y_j}^{(v)}}\right]\right\}$$

and if $D_v = \sqrt{\sum_{i=1}^{N_1}\sum_{j=1}^{N_1} p_{1_i} p_{1_j} f_{1_i 1_j} + \sum_{i=1}^{N_2}\sum_{j=1}^{N_2} p_{2_i} p_{2_j} f_{2_i 2_j} - 2\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} p_{1_i} p_{2_j} f_{1_i 2_j}}$,

then the distance is calculated as: $D = [0.3 \times abs(s_1 - s_2) + 0.7] \times D_v$.

2.1.2. Scalable Color descriptor

This descriptor performs color histogram in HSV color space encoded by a Haar transform [MPEG, 2002]. The extraction is done by quantizing the image into a 256 bin HSV color space histogram and then using the Haar transform to reduce the number of bins.

The output of the method is *a* vector with integer components, presented by a histogram with 64, 32 or 16 bins.

The distance matching can be done either in the Haar coefficient domain or in the histogram domain. In the case where only the coefficient signs are retained, the matching can be done efficiently in the Haar coefficient domain by calculating the Hamming distance as the number of bit positions at which the binary bits are different using an XOR operation on the two descriptors to be compared. This induces only a marginal loss in similarity matching precision compared to reconstructing the color histogram and performing histogram matching, while the computational cost is considerably lower.

2.1.3. Color layout descriptor

This descriptor performs spatial distribution of colors [Kasutani, 2001]. The extraction is being done as follows: the image is divided into 8x8 blocks. For each block, a single dominant color is selected. The resulting 8x8 image is then transformed into a series of coefficients using dominant color descriptors transformation. These are finally quantized to fit an assigned number of bits.

The method output is a vector with integer components, describing {*DY, DCr, DCb}* coefficients, where *Y* is the coefficient value for luminance, *Cr, Cb* coefficient values for chrominance.

For matching two descriptions *{DY, DCr, DCb}* and *{DY', DCr', DCb'}* the following formula:

$D = \sqrt{\sum_i w_{yi}(DY_i - DY_i^{'})^2} + \sqrt{\sum_i w_{bi}(DCb_i - DCb_i^{'})^2} + \sqrt{\sum_i w_{ri}(DCr_i - DCr_i^{'})^2}$ is used, where *i* represents the zigzag- scanning order of the coefficients.

2.1.4. Color structure descriptor

This descriptor is a generalization of the color histogram that encodes information about the spatial structure of the colors in an image as well as their frequency of occurrence [Messing, 2001]. The histogram is extracted in HMMD color space and non-uniformly quantizing is performed over the histogram values. This descriptor specifies spatial distribution of colors. It is calculated by letting a structuring element with image samples to visit each position in the image and then summarize the frequency of color occurrences in each structuring element location in a histogram. The structuring element always has dimensions 8x8, but the distance between the samples in the original image differs with the resolution.

The output of the method is a vector with integer components, presented by a 256 bin histogram.

The matching is done by minimizing the distance calculated as the sum of the differences between the corresponding bins in any two color-structure histograms.

2.1.5. Group-of-frame or Group-of-picture descriptor

This descriptor is a compound descriptor that expresses the color features of a collection of images or video frames by means of the scalable color descriptor [Ferman, 2000]. During the extraction the average, median or intersection scalable color histogram of the frame/picture group is calculated from scalable color histograms of each group/picture. The intersection histogram is a histogram with the minimum value for each bin over all histograms in the group.

The output of the descriptor is a vector with integer components, as in the case of scalable color descriptor.

The matching is done in the same way as for the scalable color descriptor.

2.2. Texture descriptors

The image texture is one of the most important image characteristic in both human and computer image analysis and object recognition [Manjunath, 2001]. Visual texture is a property of a region in an image. There are two texture descriptors in MPEG-7: a homogeneous texture descriptor, and edge histogram descriptor. Both of these descriptors support search and retrieval based on content descriptions.

2.2.1. Homogeneous Texture

This descriptor is aimed at texture-based image-to-image matching [Ro, 2001]. During the extraction, the mean and standard deviation of the image pixel intensities is computed. Energy and energy deviation feature values are computed by applying 30 Gabor filters in the frequency domain. The polar form used in the frequency domain in this approach is more suited for rotation invariant analysis than the Cartesian form.

The output of the method is: the average value (an integer number in the interval [0,255]); standard deviation (an integer number in the interval [0,255]); energy (30 integer numbers in the interval [0,255]); energy deviation (30 integer numbers in the interval [0,255]).

The matching is done by summing the normalized weighted absolute difference between two sets of feature vectors not using rotation or scale invariant algorithms.

2.2.2. Edge histogram descriptor

This descriptor is a texture descriptor and describes the spatial distribution of four directional edges and one nondirectional edge in three different levels of localization in an image [Park, 2000]. The localization levels are the global, the semi-global and the local level. During the extraction, the image is partitioned into 16 non-overlapping sub-images with sizes depending on the original image size. It is also divided into a preferred number of image-blocks. For each image-block, a horizontal, a vertical, a 45 degree diagonal, a 135 degree diagonal and a nondirected edge value is calculated using edge extraction filters applied on the average brightness values in four sub-blocks. If the maximum edge value is greater than a threshold value, the image-block is considered to contain the corresponding edge. Otherwise, the image-block is considered to contain no edge. The image-block edge composition in the sub-images forms a local edge histogram with a total of 80 bins (5 types of edges, for each of the16 sub-images). The global edge histogram summarizes the distribution of the different edges in the whole image by adding the corresponding local edge histogram bins into five global histogram bins one for each type of edge. The semi-global edge histogram is generated by accumulating the edge compositions in the sub-image clusters.

The output is a vector of 80 integer numbers between [0, 7].

Distance is calculated as added weighted difference between the local, global, and semi-global edge histograms respectively. Significance is measure by is the sum of absolute difference of 150 coefficients extracted from the 80 bins.

2.3. Shape descriptors

MPEG-7 supports region and contour shape descriptors. Object shape features are very powerful when used in similarity retrieval.

2.3.1. Region Shape

In the region shape descriptor, the shape of an object can be a single or multiple regions with or without holes [Kim, 1999]. The feature extraction is based on a set of Angular Radial Transform (ART) coefficients. ART is a complex 2-D transform defined on a unit disc with polar coordinates. In practice, the needed values of the basic functions are pre-calculated and put into a lookup table during the first step of the extraction. The ART transformation is then done by summing up the multiplication for each image pixel with each corresponding pixel in the lookup table, calculating the magnitudes.

The output is a vector of 35 integer numbers in the interval [0, 15].

The matching is done by calculating the minimum distance between the feature vectors for any shapes of two images. The distance for two vectors is the sum of absolute difference of coefficients.

2.3.2. Contour Shape

The contour shape descriptor presents a closed 2-D object or region contour in an image or video sequence [Mokhtarian, 1992]. During the extraction, N equidistant points are selected on the contour, starting from an arbitrary point on the contour and following the contour clockwise. The contour is then smoothed by repetitive low-pass filtering of the *x* and *y* coordinates of the selected contour points. The smoothing flattens out the concave parts of the contour. Points separating concave and convex parts of the contour and peaks in between are then identified and the normalized values are saved in the descriptor.

2.4. An example of MPEG- 7 descriptors representation

An example of MPEG-7 XML form for some descriptors on the sample image taken from TREC2002-FeatureDevelopment-mpeg1VideoSet [Smeaton, 2002], shown in Figure 1. is given after the figure.

Figure 1. A sample image from the movie “San Francisco, 1944”, KeyFrame from 1130-1407.jpg

```
<?xml version="1.0" encoding="ISO-8859-1"?><Mpeg7
xmlns="urn:mpeg:mpeg7:schema:2001"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:mpeg7="urn:mpeg:mpeg7:schema:2001"
```

```
xmlns:xml="http://www.w3.org/XML/1998/namespace"
xsi:schemaLocation="urn:mpeg:mpeg7:schema:2001 Mpeg7-2001.xsd">
<Description xsi:type="ContentEntityType"><MultimediaContent
xsi:type="ImageType">
<Image>
<MediaLocator>
<MediaUri>urn:milos:image_jpeg:SanFranc1940_KeyFrame_1130_1407</MediaUri>
</MediaLocator>
<VisualDescriptor xsi:type="ScalableColorType" numOfBitplanesDiscarded="0"
numOfCoeff="64">
<Coeff>-40 1 -12 63 -14 11 3 5 31 26 -5 9 -54 -2 12 9 -7 2 3 1 -3 5 2 -1 9
-2 1 1 -3 5 1 -4 3 3 1 3 3 2 -2 2 1 1 1 3 1 0 3 5 -9 3 2 0 -2 0 1 -3 0 0 0
-2 1 0 -3 -3</Coeff>
</VisualDescriptor>
<VisualDescriptor xsi:type="ColorLayoutType">
<YDCCoeff>48</YDCCoeff>
<CbDCCoeff>25</CbDCCoeff>
<CrDCCoeff>34</CrDCCoeff>
<YACCoeff5>15 26 22 16 14 </YACCoeff5>
<CbACCoeff2>18 17 </CbACCoeff2>
<CrACCoeff2>15 14 </CrACCoeff2>
</VisualDescriptor>
<VisualDescriptor xsi:type="ColorStructureType" colorQuant="2"><Values>0 0
0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 34 85 170 103 0 0 0 0 7 5 0 0 0
3 2 0 43 110 146 150 0 7 29 28 32 52 32 4 54 64 45 18 8 28 53 139 137 93 44
11</Values>
</VisualDescriptor>
<VisualDescriptor
xsi:type="DominantColorType"><SpatialCoherency>23</SpatialCoherency><Value>
<Percentage>1</Percentage><Index>45 38 44</Index><ColorVariance>1 0
0</ColorVariance></Value><Value><Percentage>10</Percentage><Index>191 186
163</Index><ColorVariance>0 0
0</ColorVariance></Value><Value><Percentage>9</Percentage><Index>229 230
218</Index><ColorVariance>0 0
0</ColorVariance></Value><Value><Percentage>3</Percentage><Index>111 108
111</Index><ColorVariance>0 0
0</ColorVariance></Value><Value><Percentage>4</Percentage><Index>158 154
147</Index><ColorVariance>0 0
0</ColorVariance></Value><Value><Percentage>1</Percentage><Index>117 100
84</Index><ColorVariance>0 0 0</ColorVariance></Value></VisualDescriptor>
<VisualDescriptor xsi:type="EdgeHistogramType"><BinCounts>0 1 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0 0 0 0 0 7 3 3 3 0 7 2 2 3 0 7 6 3 2 0 6 3 0 3 0 6 2 4 3 3 4
3 1 3 0 6 2 2 3 0 4 4 4 2 3 6 2 4 4 0 1 3 1 2 1 0 0 0 0 2 3 1 6
4</BinCounts>
</VisualDescriptor>
<VisualDescriptor xsi:type="HomogeneousTextureType">
<Average>188</Average><StandardDeviation>114</StandardDeviation><Energy>220
214 205 160 206 218 212 181 180 146 173 183 179 161 137 142 135 163 152 138
128 142 133 133 135 95 74 89 76 87</Energy><EnergyDeviation>223 215 205 156
207 218 202 173 180 140 164 176 169 150 122 127 122 158 143 114 126 123 129
132 129 76 58 77 53 76</EnergyDeviation></VisualDescriptor>
</Image>
</MultimediaContent>
</Description></Mpeg7>
```

3. The Use of MPEG-7 Descriptors

There are several problems, which have to be solved before evaluating the quality of different descriptors. The first problem is: how to choose the benchmark database? There is no common database used for content based benchmarking. Many researchers use the Corel image database (http://www.corel.com/). Another possibility is the collection used in MPEG-7 [MPEG, 1998], but it is also copyrighted as Corel database. Other possibilities are the databases on:

- http://elib.cs.berkeley.edu/photos/tarlist.txt,
- http://www.cs.washington.edu/research/imagedatabase/groundtruth,
- http://www.white.media.mit.edu/vismod/imagery/VisionTexture/vistex.html.

The second problem is how to measure the performance of the different descriptors. This mean to find a set of features which adequately encodes the characteristics that we intend to measure and a suitable metric. Which is the best similarity function? In 1977 Amos Tversky proposed his famous feature contrast model [Tversky, 1977]. He uses a set of binary features. In [Eidenberger, 2003] mean and standard deviation, distribution analysis and cluster analysis are used. Some of the results are: Color Layout performs badly on monochrome data. Like Color Layout, Color Structure performs inferior on monochrome data. The Dominant Color identifier performs equally well on any type of media. Scalable Color performs exactly like Color Layout and Color Structure. All color descriptors works excellent on photos but three of four perform badly on artificial media objects with few color gradations and very badly on monochrome content. An exception is the Dominant Color descriptor. This descriptor works well on each type of content. Edge Histogram performs excellent on any type of media. The Homogeneous Texture descriptor works acceptably on the Brodatz dataset. A combination of different descriptors is needed. The best descriptors for using combinations are Color Layout, Dominant Color, Edge Histogram and Texture Browsing. The others are highly dependent on these. The color histograms (Color Structure and Scalable Color) perform badly on monochrome input. Therefore, Dominant Color should be used for GoF/GoP color instead of Scalable Color. Generally, all descriptors are highly redundant and applying complexity reduction transformations could save up to 80% of storage and transmission capacity.

In [Stanchev, 2004] we generalized this result. We propose a technique for evaluating the effectiveness of MPEG-7 image features on specific image data sets, based on well defined statistical characteristics of the data set. The aim is to improve the effectiveness of the image retrieval process based on the computed similarity on these features. We also validate this method with extensive experiments with real users.

Finally, some aspects of images are captured by none of the descriptors and existing descriptors should be either refined or new visual descriptors should be added to the standard.

Conclusion

Several visual descriptors exist for representing the physical content of images, for instance color histograms, textures, shapes, regions, etc. Depending on the specific characteristics of a data set, some features can be more effective than others when performing similarity search. For instance, descriptors based on color representation might be effective with a data set containing mainly black and white images. Techniques based on statistical analysis of the data set and queries are useful.

It seems that the most intelligent descriptors are the one based on color layout. Not only does it compare the colors, but also where in the image they occur. This can be of great use if you are looking for a sunset, a face, a certain kind of landscape view etc, where similar colors usually occur in the same regions of the images. The texture and shape based search methods can also be very good, but the search results that are not among the used ground truth set can often be perceived as looking completely different compared to the query image so the use in general image databases can be questioned. On the other hand, the texture and shape based methods can recognize features such as contours and appearance that cannot be detected by the color based methods.

Even if it is not possible, in general, to overcome the semantic gap in image retrieval by feature similarity, it is still possible to increase the retrieval effectiveness by a proper choice of the image features, among those in the MPEG-7 standard, depending on the characteristics of the various image data sets (obviously, the more homogeneous the data set is, better results can be obtained).

Bibliography

[Cieplinski, 2000] Leszek Cieplinski, Results of Core Experiment CT4: Extension of Dominant Colour Descriptor, MPEG-7 TR #13-06, January 2000

[Eidenberger, 2003] H. Eidenberger, "How good are the visual MPEG-7 features?", SPIE & IEEE Visual Communications and Image Processing Conference, Lugano, Switzerland, 2003

[Ferman, 2000] A. Ferman et al., "Group-of-frame/picture color histogram descriptors for multimedia applications", Proceedings of the Storage and Retrieval of the IEEE International Conference on Image Processing", Vol. 1, Vancouver, Canada, 2000, 65-68

[Grosky, 2001] Grosky W., Stanchev P., "Object-Oriented Image Database Model", 16th International Conference on Computers and Their Applications (CATA-2001), March 28-30, 2001, Seattle, Washington (94-97).

[Kasutani, 2001] E. Kasutani, A. Yamada, "The MPEG-7 color layout descriptor: a compact image feature description for high-speed image/video retrieval", Proceeding of International Conference on Image Processing 2001, Oct. 2001, Thessaloniki, Greece 2001

[Kim, 1999] W. Kim, Y. Kim, "A new Region-Based Shape Descriptor", MPEG-7 TR#15-01, December 1999

[Manjunath, 2001] Manjunath B., Ohm J., Vasudevan V., Yamada A., Color and Texture Descriptors, IEEE Transactions on circuits and systems for video technology, V. 11, No. 6, June 2001, 703-715

[Manjunath, 2002] B.S. Manjunath, P. Salembier, T. Sikora, "Introduction to MPEG-7", Wiley, 2002

[Messing, 2001] Dean S. Messing, Peter van Beek, James H. Errico, Using Colour and Local Spatial Information to Describe Images, MPEG-7 TR #13-07, January 2001

[Mokhtarian, 1992] F. Mokhtarian, A. Mackworth, "A theory of multiscale, curvature-based shape representation for planer curves", IEEEE Transaction on Pattern analysis and machine intelligence, 14 (8), 1992, 789-805

[MPEG, 2002] "MPEG-7 Overview (version 9)", ISO/IEC JTC1/SC29/WG11N5525

[MPEG, 1998] MPEG Requirements Group, "MPEG-7: Context and objectives (version 10 Atlantic City)," Doc. ISO/IECJTC1/SC29/WG11, International Organisation for Standardisation, 1998.

[Park, 2000] D. Park, Y. Jeon, C. Won, S. Park, "Efficient use of local edge histogram descriptor", Processing of ACM International workshop on Standards, Interoperability and Practices, Marina del Rey, CA, USA, 2000, 52-54

[Rabitti, 1989] Rabitti F., Stanchev P., "GRIM_DBMS - a GRaphical IMage DataBase System", in "*Visual Database Systems*", T. Kunii (edt.) North-Holland 1989 (415-430).

[Ro, 2001] Y. Ro, M. Kim, H. Kang, B. Maniunath, J. Kim, "MPEG-7 Homogeneous Texture descriptor", ETRI Journal 23 (2), 2001, 41-51.

[Schettini, 2001] Schettini R., Ciocca G., Zuffi S., A survey of methods for color image indexing and retrieval in image databases, in Luo R., MacDonal L., (editors) Corol Imaging Science: Exploiting Digital Media, J. Willey, 2001

[Smeaton, 2002] A. Smeaton, P. Over, The TREC-2002 Video Track report, http://www-nlpir.nist.gov/projects/t2002v/results/notebook.papers/VIDEO.OVERVIEW.pdf

[Stanchev, 1999] Stanchev P., "General Image Database Model", in Visual Information and Information systems, Huijsmans, D. Smeulders A., (etd.) Lecture Notes in Computer Science 1614, 1999 (29-36).

[Stanchev, 2004] P.Stanchev, Giuseppe Amato, Fabrizio Falchi, Claudio Gennaro, Fausto Rabitti, Pasquale Savino, "Selection of MPEG-7 Image Features for Improving Image Similarity Search on Specific Data Sets", The sevent IASTED International Conference "Computer graphics and imaging", Kauai, Hawaii, 2004

[Tversky, 1977] A. Tversky, "Features of Similarity", Philosophical review, 84/4, 327-352, 1977

Authors' Information

Peter L. Stanchev – pstanche@kettering.edu

David Green Jr. – dgreen@kettering.edu

Boyan Dimitrov – bdimitro@kettering.edu

Kettering University, Flint, MI 48504, USA