Dong-Gyu Sim · Hae-Kwang Kim

Scalable representation of vector descriptors

Published online: 9 March 2006
© Springer-Verlag 2006

Abstract In this paper, a general bit-plane representation method is presented to achieve effective scalable representation for vector descriptors. Conventional vector descriptors are represented by a sequence of scalar elements and the scalability can be achieved by changing the number of elements. The proposed method represents the vector descriptors by a sequence of bit-planes that are obtained by grouping the values at the same bit positions of all the elements. In this method, the scalability is achieved by changing the number of the bit-planes. We applied the proposed scalability feature to the MPEG-7 texture descriptor and show that the performance of the proposed representation is significantly better than that of conventional representations in terms of retrieval accuracy and stability with a coarse representation.

Keywords Indexing · Scalability · Multimedia · Descriptor · MPEG-7 · Retrieval

1 Introduction

Multimedia representation is one of the most important research domains for efficient and effective management of multimedia information [1]. For a retrieval of visual content, color, shape, texture, and motion are regarded as key features and they are generally represented with vector type descriptors as in the MPEG-7 visual standard [2–4].

However, scalability has been considered as a very important feature in audio and video coding [5, 6]. The scalability for multimedia representation would also be useful for mobile and Internet environments with limited and variable network bandwidths. Both MPEG-2 and MPEG-4 support several layered scalability techniques: SNR scalability, temporal scalability, and spatial scalability. For the scalability coding techniques, a video sequence is coded into a base layer and enhancement layers. However, the scalability coding techniques can support only a few levels of bitrate. Fine granularity scalability (FGS) was introduced in the Amendment of MPEG-4 for better scalability of video coding. This scalability can be applied to the multimedia content description. In mobile computing environments in which available resources such as communication bandwidth and computing capability are limited, the mobile device might extract and transfer a coarse descriptor to a remote retrieval system.

The retrieval system would perform a matching operation of the received coarse descriptor with full precision descriptors in its database. In the Internet environment, the granularity of scalable descriptors may be automatically adjusted to dynamic network conditions. In this paper, we propose a scalable vector description method based on the bit-plane representation used for the MPEG-4 FGS.

Vector descriptors defined in the MPEG-7 standard consist of several scalar elements of the same data type. Scalable representation can be obtained by adjusting the number of elements. However, scalability cannot be effectively achieved with the partial representation because the descriptors have not been designed for scalability. In this paper, an effective scalable bit-plane representation method is presented to achieve the scalable accessibility for vector-type multimedia information.

In the rest of paper, the proposed scalable bit-plane representation method is explained in detail and then comparative experimental results in retrieval accuracy regarding the scalability are presented. In the experiment, the homogeneous texture descriptor in the MPEG-7 Visual standard is used.

2 Scalable bit-plane vector descriptor representation

A vector descriptor consists of several scalar elements of the same data type. Conventionally, scalability can be obtained
by adjusting the number of elements. For the vector descriptor, each element is considered to be independent from other elements. The conventional representation of the MPEG-7 homogeneous texture descriptor is given by

\[
\{DC, STD, n, F, S\} = \{DC, STD, n, F_1, F_2, \ldots, F_n, S_1, S_2, \ldots, S_n\},
\]

where DC is the average intensity, STD is the standard deviation of the intensity, and \(n\) is the number of frequency channels, which is set to 30 in the standard. \(F\) and \(S\) are the feature vectors of which \(F_i\) and \(S_j\) are the first and the second order moments of the \(i\)th frequency channel and are represented with 8-bit precision. In this descriptor, each element has its own importance and is independent from the others. For the scalability, the high frequency elements of the descriptor may be removed for a coarse representation because human perception focuses more on lower frequency information. Conventional vector descriptors are represented by a sequence of scale elements. The scalability can be achieved by changing the number of elements. However, for many vector descriptors, each vector descriptor has its own rule of importance that organizes its elements. Furthermore, it would be hard to prioritize the importance of individual elements. The scalable representation for the vector descriptors may not yield the best performance with a partial number of scalar elements.

The proposed scalable method represents a vector descriptor with a sequence of bit-planes in which each bit-plane is built by gathering the values at the same bit position of all the elements. The bit-planes are sequenced in the order of the most significant bit (MSB) first. In the proposed method, a coarse representation is achieved by removing several of the least significant bit-planes, resulting in a reduced number of bit-planes. For example, a general vector descriptor is denoted by

\[
X = \{X_1, X_2, \ldots, X_N\},
\]

where \(N\) represents the number of elements and \(X_n\) represents the value of the \(n\)th element expressed by

\[
X_n = \sum_{m=1}^{M} X_{n,m} 2^{M-m},
\]

where \(X_{n,m}\) is the binary value of the \(m\)th bit of \(X_n\) and \(M\) is the number of bits for an element. \(X_n\) can be represented by a binary sequence \(\{X_{n,1}, X_{n,2}, \ldots, X_{n,M}\}\). The proposed bit-plane representation is expressed by

\[
BS_M(X) = \{Y_1, Y_2, \ldots, Y_M\},
\]

where \(Y_m\) denotes \(\{X_{1,m}, X_{2,m}, \ldots, X_{N,m}\}\) which is a sequence of the binary values at the \(m\)th bit position of \(N\) elements and \(N\) is the total number of values for a vector descriptor.

With this representation, the matching distance between a query descriptor, \(Q(m_1) = \{Y^Q_1, Y^Q_2, \ldots, Y^Q_{m_1}\}\) with \(m_1\)-bit precision and a database descriptor \(D(m_2) = \{Y^D_1, Y^D_2, \ldots, Y^D_{m_2}\}\) with \(m_2\)-bit precision is calculated with the following modified L1 measure:

\[
\text{Dist}(Q(m_1), D(m_2)) = \sum_{m=1}^{\min(m_1, m_2)} d(Y^Q_m, Y^D_m) \times 2^{M-m},
\]

where

\[
d(Y^Q_m, Y^D_m) = \sum_{n=1}^{N} (X^Q_{n,m} \oplus X^D_{n,m}) \oplus
\]

is the logical exclusive-or operator and \(N\) is the number of elements of the vector. The L1 measure is calculated with only the most significant common bit-planes between two given descriptors. The MPEG-7 homogeneous texture descriptor is represented based on the proposed method as

\[
\{DC, STD, m, BS_M\{F\}, BS_M\{S\}\} : M = 8.
\]

### 3 Experimental results

The effectiveness of the proposed algorithm is shown with the comparison of the retrieval accuracies between a conventional scalable representation and the proposed scalable bit-plane representation. The comparison was conducted with the MPEG-7 homogeneous texture descriptor using two well-known texture datasets. The first dataset consists of 109 Brodatz texture images (512 × 512) [7] and 7 USC (512 × 512) texture images. 109 Brodatz texture images were obtained by digitizing the Brodatz photo album. Figure 1 shows six examples of textured surfaces. The second dataset from Information and Communications University (ICU) was constructed by taking real textures of outdoor objects with a digital camera. The dataset consists of 52 texture images (512 × 512). Both datasets contain real-world texture. Each image is divided into 16 texture blocks (128 × 128). The 16 texture blocks are considered as being in a texture class. For a query image out of the 16 block images, our purpose is to find the remaining 15 texture blocks coming from the same original image. The retrieval rate (RR) for the \(j\)th query image of the \(i\)th texture class is defined by

\[
RR_{ij} = \frac{\text{number of relevant images retrieved}}{\text{from the 15 ground truth images}}
\]

where \(i\) represents the texture index in the first and second dataset. The texture index ranges from 1 to 109 + 7 = 116 and from 1 to 52 for the first and second datasets, respectively. The index of the query image for the \(i\)th texture is \(j (1 \leq j \leq 16)\). The average class retrieval rate (ACRR) for the \(i\)th texture class is defined by

\[
\text{ACRR}_i = \sum_{j=1}^{J} RR_{ij}/J
\]
Fig. 1 Six sample textures from the Brodatz album
where $J$ is equal to 16 for this dataset. As a result, the total average retrieval rate (TARR) is denoted by

$$TARR = \frac{1}{I} \sum_{i=1}^{I} \text{ACRR}_i$$

where $I$ is equal to 116 for this dataset.

The comparison is performed with the proposed and conventional homogeneous texture descriptors. At first, for each texture image, its MPEG-7 homogeneous texture descriptor is extracted in the conventional representation form. Then, the descriptor in the conventional form is converted to the proposed bit-plane representation. For the conventional representation, scalability is achieved by discarding high frequency elements out of all the conventional elements. For the bit-plane representation, descriptors of different scales are obtained by taking several most significant bit-planes from all the bit-planes. L1 distance measure is used for retrieving both representations. Note that we made use of 30 frequency channels in this experiment.

Figure 2a shows the average retrieval accuracy curves of both representations for the first dataset. The horizontal axis represents the number of bits for the scaled descriptors. The curve clearly shows that the proposed bit-plane representation yields much better than the conventional representation in terms of retrieval accuracy with the same amount of bits, especially at the low bit region. For the conventional method, ACRR is only 26.8% with 32-bit representation. The ACRR of the proposed algorithm is 42.48% with a 30-bit representation. Figure 2b shows Normalized Standard Deviations of Retrieval Accuracy (NSDRA) curves for both representations. NSDRA is defined as the standard deviation of retrieval accuracies over all texture classes (116 classes for the first dataset and 52 classes for the second dataset), normalized by the average retrieval accuracy. NSDRA is determined by NSDRA = STD(ACRR)/TARR. A lower value NSDRA represents better stability of the retrieval accuracy. The curves show that the proposed representation is more stable than the conventional one with a small amount of bit information in terms of retrieval accuracy. We can achieve more stable and better performance regardless of texture characteristics.
Figure 3a shows the average retrieval accuracy curves from both representations for the ICU dataset. The proposed description is significantly better than the conventional representation because the proposed scalable description makes use of all the frequency components represented with a limited amount of bits. The conventional representation method yields an ACRR of 34.7% with 32-bit representation while the proposed method yields an ACRR of 50.1% with only 30-bit representation. The ACRR for the ICU dataset is higher than that of Brodatz set because the Brodatz set has many similar texture classes. Additionally, the variation of the retrieval accuracy by the proposed representation is smaller than that by the conventional one as shown in Fig. 3b, meaning that the proposed description is more stable and is likely to be effectively applied to other datasets.

4 Conclusion

In this paper, a new generic scalable bit-plane representation is presented for vector descriptors. Experimental results with the MPEG-7 homogeneous texture show that the proposed representation method is more effective and more stable than the conventional representation in terms of retrieval accuracy.

References
