Journal of Korea Multimedia Society Vol. 12, No. 6, JUNE 2009(pp. 785-793)

Video Sequence Matching Using Normalized Dominant Singular Values

Kwang-Min Jeong[†], Joon-Jae Lee[††]

ABSTRACT

This paper proposes a signature using dominant singular values for video sequence matching. By considering the input image as matrix A, a partition procedure is first performed to separate the matrix into non-overlapping sub-images of a fixed size. The SVD(Singular Value Decomposition) process decomposes matrix A into a singular value-singular vector factorization. As a result, singular values are obtained for each sub-image, then k dominant singular values which are sufficient to discriminate between different images and are robust to image size variation, are chosen and normalized as the signature for each block in an image frame for matching between the reference video clip and the query one. Experimental results show that the proposed video signature has a better performance than ordinal signature in ROC curve.

Key words: content-based video matching, video sequence matching, singular value decomposition, ordinal measure

1. INTRODUCTION

With the ever increasing amount of digital media (images, audio, and video) available on the Web and in multimedia databases, the need for content-based copy detection schemes has also become evident for handling digital contents and protecting intellectual property rights.

Currently, the most widely used technique for content-based copy detection is a sequence matching approach, where multiple sequence frames are used as the basis for matching, as opposed to matching single video frames. Features like intensity rankings and color histograms are extracted from the original video frames to create reference signatures in a database. The same features, extracted from the query video sequence are then matched to the reference signatures to determine if the query video sequence is a copy of the original[1,2]. Related work includes the following:

Jain et al.[3] proposed a sequence matching method based on a set of key frames. Although motion information is included with the key frames, it is not clear. Meanwhile, Naphade et al.[4] proposed an algorithm for matching video clips that uses the histogram intersection of YUV histograms of the DC sequence of an MPEG video. However, this technique does not evaluate variations between copies, such as signal modifications or display format conversions. Mohan[1] used the ordinal measure originally proposed by Bhat et al.[5] for video sequence matching, then Hampapur et al.[6] compared the ordinal measure technique with techniques using a motion signature and color signature, and showed that matching based on an ordinal signature produced the best performance. The ordinal measure is also insensitive to intensity

※ Corresponding Author : Joon-Jae Lee, Address : (705-701) Daemyeong 3-dong, Nam-gu, Daegu, Korea, TEL : +82-53-620-2177, FAX : +82-53-620-2198, E-mail : joonlee@kmu.ac.kr

Receipt date : Oct. 29, 2008, Approval date : Mar. 4, 2009

[†] Information Communication Subdivision, Kyungnam College University of Information & Technology (E-mail : kmjeong@kit.ac.kr)

[††] Dept. of Game Mobile Contents, Keimyung Univ.

value changes and has a low memory requirement for storing/indexing the signature. For this reason, C. Kim[7] proposed ordinal measure using DCT coefficient.

Our video copy detection system uses the singular value decomposition (SVD) which is known for its capabilities of deriving the low dimensional refined feature space from a high dimensional raw feature space, and capturing the essential structure of a data set in the feature set. The SVD process decomposes matrix A into a singular value-singular vector factorization, where the singular values represent the energy of matrix A projected on each subspace. SVD has an excellent energy compaction property and the large singular values represent the dominant information in matrix A. Thus, when considering the input image as matrix A, the singular values and their distribution represent useful information on the contents of A. From the simple normalization which took all of singular values to make normalized signatures, we got a good image signature[8]. This image signature shows good results for video sequence matching. But this signature did not solve a problem according to resolution change.

The number of singular values differs according to the image size. In other words, if the size of an image change, the number of singular vectors also changes along with the number of singular values and their distribution. However, it is important to obtain the same number of features as ordinal methods have the same number of features irrespective of a variation in the image size.

Unlike the simple normalization used in the above methods to obtain features invariant to image size variation, it is not suitable to apply the same normalization method to the proposed features using singular values. However, since the energy distribution of the singular values is concentrated on just a few dominant singular values, equivalent features can be obtained by normalizing an appropriate number of these dominant values. As a result, this solves the vector size problem and increases the computational efficiency. Yet, the problem is how to choose the appropriate number of singular values to maximize the matching accuracy. In this paper, this is achieved by choosing k singular values including a half of energy of singular values.

The similarity between two video clips can be treated as the similarity between the shapes represented by the singular values of two corresponding hyper-ellipses, if the orientation is not considered. Therefore, the distance or dissimilarity metric between image frames can be represented by the Euclidean distance of the singular values. In experimental results, the matching performance is demonstrated in comparison with the ordinal measure.

The rest of the paper is organized as follows. Section 2 describes the properties of singular value as image signature and proposed image signature, then Section 3 provides a detailed explanation of the video sequence matching algorithm using the proposed feature. Experimental results are presented in Section 4, and some final conclusions are given in Section 5.

2. SIGNATURE EXTRACTION

2.1 Image representation of singular value decomposition (SVD)

Let an image of size $M\times N$ be image matrix A of dimensions $M\times N$, where $M\geq N$. It is possible to represent this image in the r-dimensional subspace, where r is the rank of A, and $r\leq N$. Singular value decomposition is then a factorization of matrix A into orthogonal matrices.

$$A = USV^T \tag{1}$$

where U is an $M\times r$ matrix and consists of the orthonomalized singular vectors of A^TA, and S is an $r\times r$ diagonal matrix consisting of the 'singular values' of A, which are the non-negative square roots of the singular values of A^TA. These singular values, denoted by σ_i, i=1, 2, ⋯ , r, are sorted in a non-increasing order, i.e.

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0 \qquad (2)$$

An important property of U and V is that they are mutually orthogonal. The singular values (σ_i) represent the importance of individual singular vectors in the composition of the matrix. In other words, the singular vectors corresponding to large singular values include more information about the matrix than the other singular vectors.

2.2 Singular value as image signature

The dominant singular vectors of image matrix A represent dominant information of A and their magnitude are singular values. The singular values represent the energy of matrix A projected on each subspace, where the singular values and their distribution carry useful information about the contents of A, and can vary drastically from image to image[8]. For most images, only a very few larger singular values dominate, while all the other singular values are quite small. The four sample images and their corresponding 10-dominant singular values are shown in Figs. 1 and 2. The dis tributions of the singular values for the images are quite different.

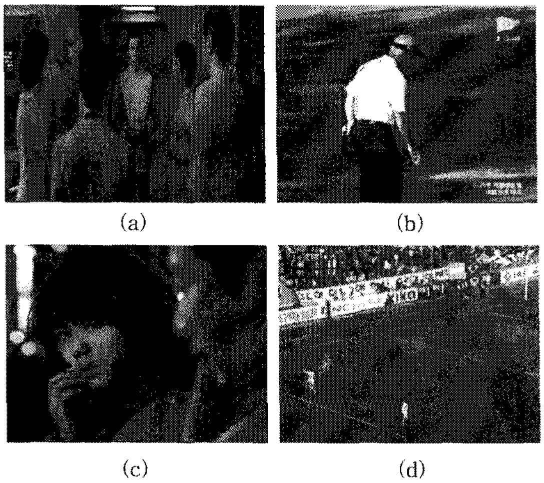

(a) (b) (c) (d)

Fig. 1. The example of four different images; (a) movie, (b) golf, (c) drama, (d) football

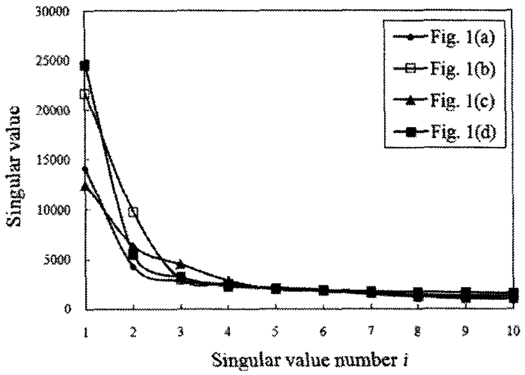


Fig. 2. The 10-dominant singular values for Fig. 1 sample images

One of the important characteristics for an image signature is robustness to media transform such as image size variation and compression format change. Therefore, the singular values need to undergo a normalization process. If an image of size $M \times N$ has an image matrix A of dimensions $M \times N$, where $M \geq N$, this image can be represented in r-dimensional subspace, where r is the rank of A, and $r \leq N$. The previous paper[9] suggested image signature by simple normalization of the singular values as follows :

$$\bar{\sigma}_i = \frac{\sigma_i}{\sum_{j=1}^{r} \sigma_j}, \qquad i = 1, 2, \cdots, r \qquad (3)$$

where $\bar{\sigma}_i$ is the ith normalized singular value and σ_j is the jth singular value. As shown in Fig. 3, the normalized singular values as an image signature can discriminate different images. The magnitude of the largest singular value is 0.32 in Fig. 5, which means it includes 32% of the image information. These singular values are sorted in a non-increasing order using Eq. (2) so they can be used as a signature in the order of importance.

However, since this image signature is not robust to a resolution change, the dimension of video clip has to be adjusted to be same. Fig. 5 shows

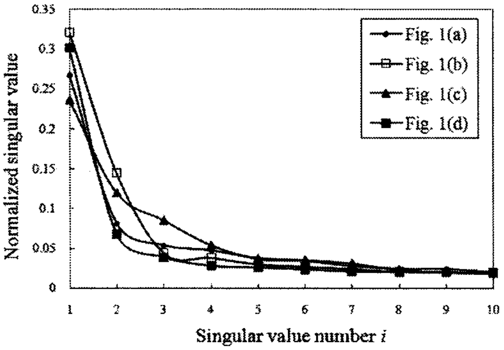
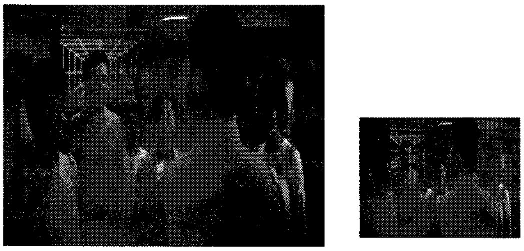


Fig. 3. The 10-dominant normalized singular values for different sample images in Fig. 1

a comparison of the normalized singular values as an image signature for the Fig. 4 images. The comparison results show a big difference in the signatures for the two images with different resolutions. This is because, if the dimensions of an image change, the number of singular vectors also changes, along with the number of singular values and their distribution.



(a) (b)

Fig. 4. Different resolution images: (a) 320×240 original image, (b) 160×120 image

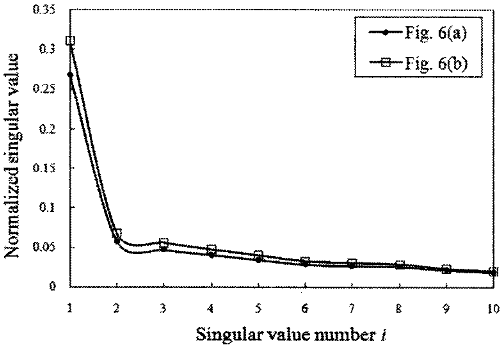


Fig. 5. 10-dominant normalized singular value distribution for different resolution image in Fig. 6

2.3 Proposed image signature

In general the large singular values represent the dominant information for image matrix A, while the small singular values include little information and can be considered as introduced by noise[10]. Therefore, if only dominant k singular values are used, the size problem occurring due to different vector numbers can be solved and the computational efficiency can also be raised. From the comparison of the singular values for four different images in Fig. 2, only the first 4 dominant singular values are different, while the others have very similar values. Thus, the proposed image signature uses only the dominant 5 singular values that include the fifth singular value as a redundancy. The truncated normalized singular values are as follows:

$$\bar{\sigma}_{T,i} = \frac{\sigma_i}{\sum_{j=1}^{5}\sigma_j}, \qquad i = 1, 2, \cdots, 5 \qquad (4)$$

where $\bar{\sigma}_{T,i}$ is the *i*th truncated normalized singular value and σ_j is the *j*th singular value. Table 1 summarizes the averaged normalized singular

Table 1. Averaged normalized singular values of the 2,000 random sampled images from database

Normalized singular value	Image size			
	320×240		160×120	
	Value	Cumulative value	Value	Cumulative value
$\bar{\sigma}_1$	0.3388	0.3388	0.3663	0.3663
$\bar{\sigma}_2$	0.0753	0.4141	0.0827	0.4490
$\bar{\sigma}_3$	0.0512	0.4653	0.0566	0.5056
$\bar{\sigma}_4$	0.0391	0.5044	0.0432	0.5488
$\bar{\sigma}_5$	0.0320	**0.5364**	0.0351	**0.5839**
$\bar{\sigma}_6$	0.0272	0.5636	0.0298	0.6137
$\bar{\sigma}_7$	0.0237	0.5873	0.0259	0.6396
$\bar{\sigma}_8$	0.0210	0.6083	0.0227	0.6623
$\bar{\sigma}_9$	0.0188	0.6271	0.0203	0.6826
$\bar{\sigma}_{10}$	0.0171	0.6442	0.0183	0.7009

values for 2,000 randomly sampled images from a database. The cumulative sum of the 5 largest dominant singular values contains above 50% of the information on the images without relation to the image size.

The proposed image signature can be calculated quickly using the power method, as only dominant 5 singular values are used, resulting in computational efficiency. In addition, the spatial information on an image frame is incorporated by dividing each frame into 3×3 blocks, and calculating the normalized singular values for each block.

The robustness of the proposed image signature to a resolution change is shown in Fig. 6. The results confirm that the proposed image signatures are very similar without relation to the image size.

2.5 Distance measure

Assuming that ν_i is one of the singular vectors V and σ_i is the corresponding singular value, then ν_i determines the orientation of one semi-axis of the hyper-ellipse and σ_i measures the length of the corresponding axis. As such, the shape and orientation of the hyper-ellipse is a description of the characteristics of the image. Fig. 7 illustrates two different hyper-ellipses in a 2-dimensional plane as an example As a result, the similarity between two video clips can be treated as the similarity between the shapes representing by singular values of two corresponding hyper-ellipses, if the orientation is not considered. Therefore, the above description leads to the following distance or dissimilarity metric between image frame A and frame A′.

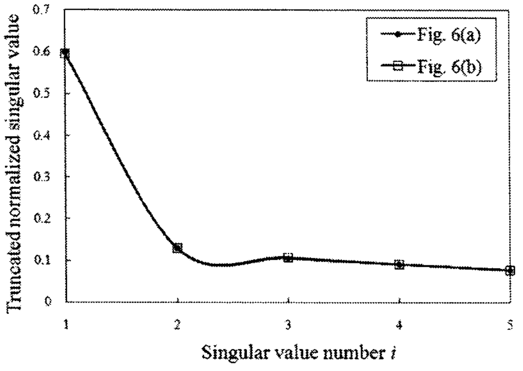


Fig. 6. Resolution variation for proposed image signature: 5-dominant proposed image signatures for two images in Fig. 4

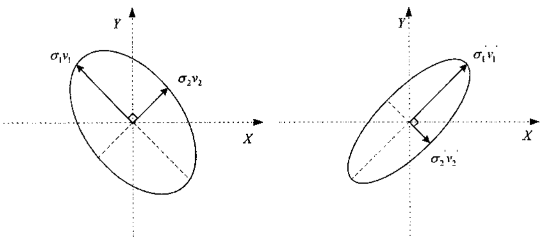


Fig. 7. Illustrations of two 2-D hyper-ellipses with two dominant feature vectors

$$Dist(\mathrm{A},\mathrm{A}') = \sum_{i=1}^{r} \left|\sigma_i - \sigma_i'\right| \qquad (5)$$

Case of normalized singular value is the same as Eq. 11.

$$Dist_{nor}(\mathrm{A},\mathrm{A}') = \sum_{i=1}^{r} \left|\overline{\sigma}_{T,i} - \overline{\sigma'}_{T,i}\right| \qquad (6)$$

3. VIDEO SEQUENCE MATCHING ALGORITHM

A video is composed of continuous image frames. Therefore, a video sequence clip with N frames is denoted by

$$C = \{C[1], C[2], \cdots, C[N]\}, \qquad (7)$$

and the ith frame with m partitions can be expressed as follows

$$C[i] = \{C^1[i], C^2[i], \cdots, C^m[i]\}, \qquad (8)$$

where C^j denotes the sequence of the jth partition.

A sub-video of C is also defined as $C[p : p+N-1]$, where the number of frames is N and the first frame is $C[p]$, $1 \le p \le n-N-1$. Let the normalized singular value matrix for the ith frame of the query video clip $C_q[i]$ be

$$\sigma_{q,1}(i), \sigma_{q,2}(i), \cdots, \sigma_{q,k}(i),\ k = 1, \cdots, r \qquad (9)$$

and if the frame has m partitions, the normalized singular value matrix for the jth partition of the ith frame $C_q^j[i]$ can be defined as

$$\sigma_{q,1}^j(i), \sigma_{q,2}^j(i), \cdots, \sigma_{q,k}^j(i),\ j=1,\cdots,m \qquad (10)$$

The singular value distance between the ith frame of the reference video sequence $C_r[p:p+N]$, $C_r(p+1)$, $1 \leq i \leq N$ and the ith frame of the query video clip $C_q[i]$ can then be defined as

$$d(C_q[i], C_r[p+i]) = \sum_{j=1}^{m}\sum_{k=1}^{r}|\sigma_{q,k}^{j'} - \sigma_{r,k}^{j'}[p+i]| \qquad (11)$$

As such, the dissimilarity between the two sequences $D(C_q, C_r[p:p+N-1])$ is computed by averaging over N dissimilarities, i.e.

$$D(C_q, C_r[p:p+N-1]) = \frac{\sum_{i=1}^{N} d(C_q[i], C_r[p+i])}{N} \qquad (12)$$

Given a query video clip C_q with N frames, the reference video sequence C_r, C_q is compared to the sub sequence $C_r[p:p+N-1]$. The detailed matching procedure is as follows.

Step 1) Set p to be 1.

Step 2) Compute the distance between the two sequences,
C_q and $C_r[p:p+N-1]$.

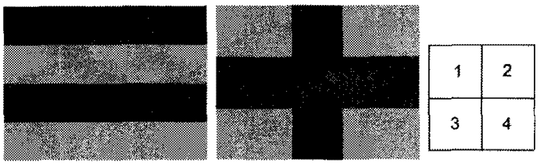
Step 3) Increase p by 1. Repeat Step 2 until $p=n-N$,
where n is the number of frames in the reference video.

Step 4) From the sequence of distances, find global minima.
The global minima point is declared as the best match location.

4. EXPERIMENTS AND DISCUSSION

4.1 Comparisons with ordinal measure characteristics for artificial images

The ordinal measure was proposed by Bhat et al. for computing image correspondence[5]. And Mohan applied the ordinal measure to video sequence matching[1]. Ordinal feature obtained by as follows. The video frame is partitioned into $N=N_x \times N_y$ equal-sized blocks and the average gray level in each block for N sub-image is computed. Then the set of average intensities is sorted in ascending order and the rank is assigned to each block using integer 1 to$N$. The main advantages of ordinal measure are low processing time, low memory space, and robustness to intensity changes. However there exist many real video or movie scenes with same ordinal feature even though they have entirely different contents or structure of image. Figs. 8(a) and (b) show examples where the ordinal feature fails in the case of artificial images that have the same average intensity value for block pixels irrespective of a different structure or content while the proposed system has a good discriminatory power due to different singular values corresponding to different structural information, as shown in Table 2.



(a) (b) (c)

Fig. 8. Artificial images for comparison of signature characteristic; (a) two horizontal lines image, (b) cross lines image, (c) sub-block number of images

Table 2. Ordinal and the proposed signature for artificial images

	Ordinal				Proposed image signature			
	Fig. 8(a)		Fig. 8(b)		Fig. 8(a)		Fig. 8(b)	
#block	mean	rank	mean	rank	$\overline{\sigma_{T,1}}$	$\overline{\sigma_{T,2}}$	$\overline{\sigma_{T,1}}$	$\overline{\sigma_{T,2}}$
1	123.65	1	123.65	1	1	0	0.9219	0.0781
2	123.65	2	123.65	2	1	0	0.9219	0.0781
3	123.65	3	123.65	3	1	0	0.9219	0.0781
4	123.65	4	123.65	4	1	0	0.9219	0.0781

4.2 Performance measure for video sequence matching

The performance of the proposed algorithm was plotted based its receiver operating characteristics (ROC) curve, which is a plot of the false positive rate (R_{FP}) versus the false negative rate (R_{FN}). Let N_T be the total number of match tests conducted, with F_N the number of false negatives (clips that should have been matched, yet were not) and F_P the number of false positives (clips that matched but were not part of the reference set, as another video was used for the R_{FN}). Thus, the R_{FN} and R_{FP} were as follows:

$$R_{FN}(\tau)=\frac{F_N}{N_T},\ R_{FP}(\tau)=\frac{F_P}{N_T} \qquad (13)$$

where τ is the normalized threshold value varying from zero to one. The ROC curves were computed by varying τ. A good ROC curve lies very close to the axes, while an ideal curve pass through (0,0), i.e. zero R_{FN} and zero R_{FP}.

4.3 Matching results for real video sequence

An MPEG-1(320×240) video reference sequence with 200,000 frames was used that included a variety of sub sequences: movie, football, golf, CF, and TV drama. The query video sequence was derived from MPEG-1(160×120) encodings of the reference video and a home video composed of 51,000 frames.

For the test, 100 query clips with different clip lengths were sampled from the query video sequence. The ROC curves were computed when varying τ with an increment of 0.5%.

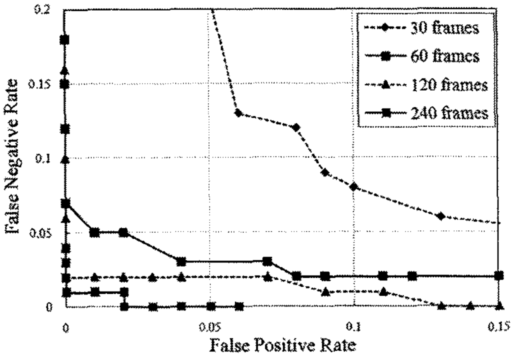


Fig. 9. ROC curves: ordinal signature for 3×3 partitions

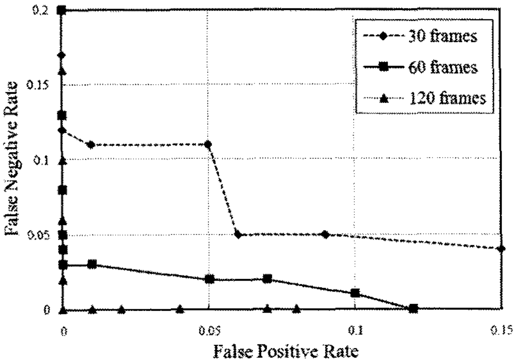


Fig. 10. ROC curves: 1-dominant proposed image signature for 3×3 partitions.

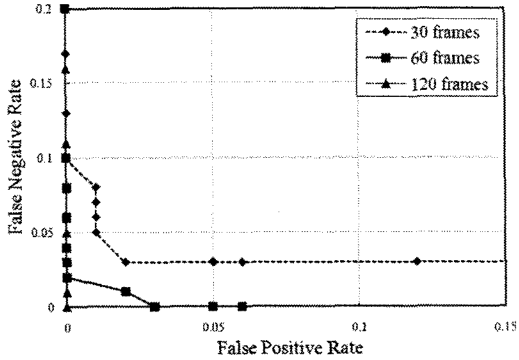


Fig. 11. ROC curves: 2-dominant proposed image signature for 3×3 partitions

The ROC curves for ordinal matching and the proposed method using one dominant signature are shown in Figs. 9 and 10, where one ROC curve is given for each clip length. As expected, the matching performance improved when increasing the clip length. Based on the results, the proposed image signature produced a better performance than the ordinal signature, as the proposed signature overcomes a weakness of the ordinal signature when the block mean is the same but the contents are different. In addition to producing the best performance for very short clips, if the proposed image signature is calculated using a fast algorithm, it can be computed in real-time processing for live video.

The ROC curves for the proposed image signature using 2 dominant signatures is shown in Fig. 11. From the results in Fig. 10 and Fig. 11, the more signatures used, the higher the matching efficiency, although more time is required.

4.4 Memory requirement

The proposed image signature is composed of decimal fraction from zero to one. When proposed image signature use only one dominant singular value for video matching process, memory requirement of the signature is 18 bytes/frame (2 bytes/block). So signatures convert to integer as follows

$$\overline{\sigma'_{T,i}} = roundoff\ (255 \times \overline{\sigma_{T,i}}) \qquad (14)$$

where $0 \le \overline{\sigma_{T,i}} \le 1$ and $0 \le \overline{\sigma'_{T,i}} \le 255$. From Eq. (14), 2 bytes decimal fraction is converted to 1 byte integer. Therefore the memory requirement is reduced to 9 bytes/frame.

5. CONCLUSION

The dominant singular values and their distribution based on singular value decomposition, which decomposes an image into a singular value-singular vector factorization, were proposed as an image signature for video sequence matching. To obtain features that are robust to image size variation, the dominant 5 singular values, corresponding to half the accumulated energy and sufficient to discriminate between different images, are chosen and normalized. As a result, the proposed algorithm can reduce the noise effect following a media transform and increase the computational efficiency using a fast algorithm. In addition, to accelerate the SVD computation and enhance the accuracy through the use of spatial information from an image, the input image is partitioned into 9 non-overlapping sub-images.

In experiments, the proposed image signature was evaluated in comparison with the ordinal measure. The results demonstrated that the proposed image signature produced a better performance than an ordinal signature. Additionally, if only one dominant proposed image signature is used for each image block, the memory required for storing/indexing the signatures can be minimized to the same as an ordinal signature with 9 bytes/frame. Another advantage of the proposed signature is that multiple signatures based on the magnitude of the singular values can be used to achieve a higher matching efficiency.

REFERENCES

[1] R. Mohan, "Video sequence matching," *Proceedings of International Conference on Audio, Speech and Signal Processing,* Vol.6, pp. 3697-3700, Jan. 1998.

[2] A. Hampapur and R. M. Bolle, "Comparison of distance measures for video copy detection," *Proceedings of the IEEE International Conference on Multimedia and Expo,* pp. 22-25, Aug. 2001.

[3] A. K. Jain, A. Vailaya, and W. Xiong, "Query by video clip," *Multimedia Systems,* Vol.7, No.5, pp. 369-384, 1999.

[4] M. Naphade, M. Yeung and B. Yeo, "A novel scheme for fast and efficient video sequence matching using compact signatures," *Proceedings of SPIE Storage and Retrieval for Media Database 2000,* Vol.3972, pp. 564-572, Jan. 2000.

[5] D. Bhat and S. Nayar, "Ordinal measures for image correspondence," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol.20, Issue 4, pp. 415-423, Apr. 1998.

[6] A. Hampapur, K. H. Hyun and R. M. Bolle, "Comparison of Sequence Matching

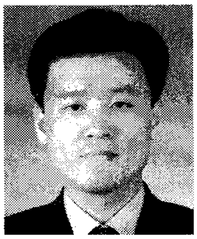
Techniques for Video Copy Detection.," *Proceedings of SPIE, Storage and Retrieval for Media Database 2002,* Vol.4676, pp. 194-201, Jan. 2002.

[7] C. I. Kim, "Content-based image copy detection," *Signal Processing : Image Communication,* Vol.18, No.3, pp. 169-184, Mar. 2003.

[8] C. J. Lu and D. M. Tsai, "Defect inspection of patterned thin film transistor-liquid crystal display panels using a fast sub-image-based singular value decomposition," *International Journal of Production Research,* Vol.42, No.20, pp. 4331-4351, Oct. 2004.

[9] K. M. Jeong, J. J. Lee, and Y. H. Ha, "Video Sequence Matching Using Singular Value Decomposition", *ICIAR 2006,* LNCS4141, pp. 426-435, Sept. 2006.

[10] J. Gu, L. Lu, R. Cai, H. J. Zhang, and J. Yang, "Dominant Feature Vectors Based Audio Similarity Measure", *Proceedings of 5th Pacific Rim Conference on Multimedia 2004,* LNCS 3332, pp. 890-897, Nov. 2004.



Kwang-Min Jeong

1989. Kyungpook National University (BS)
1991. Kyungpook National University (MS)
2006. Kyungpook National University (Ph.D)
1991.~1998. DAERYUNG IND. INC. Research Manager
1998.~Kyungnam College University of Information & Technology Associate Professor
Areas of Interest : Image Processing, Digital Signal Processing, and Computer Vision

Joon-Jae Lee

1986. 8 Kyungpook National University (BS)
1990. 8 Kyunpook National University (MS)
1994. 8 Kyunpook National University (Ph.D)
1998. 3~1999. 2 Georgia Institute of Technology. Visiting Professor
2000. 3~2001. 2 PARMI Corporation. Research Manager
1995. 3~2007. 8 Dongseo University, Associate Professor
2007. 9~Keimyung University, Associate Professor
Areas of Interest : Image Processing, 3-D Computer Vision, and Computer Graphics