Qualitative Comparison of Audio and Visual Descriptors Distributions

Stanislav Barton*, Valerie Gouet-Brunet*, Marta Rukoz†, Christophe Charbuillet‡ and Geoffroy Peeters‡
*CNAM/CEDRIC, 292, rue Saint-Martin, F75141 Paris Cedex 03
†LAMSADE CNRS UMR 7024, Place de Lattre de Tassigny 75775 Paris Cedex 16
‡IRCAM, 1, place Igor-Stravinsky, 75004 Paris

***Abstract*—A comparative study of distributions and properties of datasets representing public domain audio and visual content is presented. The criteria adopted in this study incorporate the analysis of the pairwise distance distribution histograms and estimation of intrinsic dimensionality. In order to better understand the results, auxiliary datasets have been also considered and analyzed. The results of this study provide a solid ground for further research using the presented datasets such as their indexability with index structures.**

I. Introduction

In order to make the multimedia data searchable by its content, various methods of mapping the multimedia content into high-dimensional spaces have been introduced for images [4] and audio [7]. Since, like all high dimensional data suffer from *the curse of dimensionality*, we would like to analyze such data to understand its nature and to give other researchers a base ground for further work, e.g., indexing. In [2] was proven that the complexity of searching the data grows exponentially with the dimensionality of data thus it is important to be able to set the tradeoff between fine grained information as high-dimensional feature vectors and *good* searchability of the data.

Therefore, in this paper we present a comparative study of the properties of multimedia datasets representing visual and audio descriptors acquired from the public domain content provided by EWA [1]. The data is investigated in terms of pair-wise distance distribution and of the estimation of the intrinsic dimensionality. Because we focus on multimedia in general, we incorporate in our study both visual and audio data. Using the same methodology and criteria and by comparing the results, we would like to depict the different characteristics of these two types of multimedia considering also the datasets where the characteristics is known.

A. Visual Descriptors

Color, texture and shape have been identified as the main low-level and global descriptors that can characterize the image content. For example, the visual features included in the MPEG-7 standard consist of histogram-based descriptors, spatial color descriptors and texture descriptors [10]. They are called *global descriptors* because they resume in one feature vector all the image content, in comparison to other *local* techniques, e.g., interest point identification, which can result in more than one feature vector per investigated image.

Global features has been used for a long time to characterize the visual aspect of images. They have the advantage of encapsulating some global semantics or ambiance such as *indoor* or *painting*, while requiring a low amount of data to describe it. Despite the simplicity, such family of descriptors was evaluated as relevant for content-based information retrieval applications [5].

In this study, color histogram as global description of the color distribution present in the image [12] is used. Such histogram counts the proportion of each color in the image. The color space chosen is classical RGB (for Red, Green and Blue). Because a 24 bytes image is able to store more than 17 millions of colors, a discretization of the space is required to reduce the number of colors to count. By considering for example 4 bits for the Red channel, 4 bits for the Green one and 8 for the Blue one, the RGB descriptor obtained is a $4 \times 4 \times 8 = 128$ feature vector. The similarity measure used is Euclidean distance – L_2.

B. Audio Descriptors

Global audio descriptors used for music similarity are mainly based on the modeling of short term audio features. We present here a study on the model proposed by [11]. The main idea of this approach is to describe the temporal evolution of a sequence of short term descriptors.

Obviously, the choice of the short term feature is fundamental. In order to provide a general audio description, we selected four different short term descriptors. The Mel Frequency Cepstrum Coefficient (MFCC) which gives a robust cepstral shape description, the Chroma descriptors which provides an harmonic representation, the Spectral Crest Factor (SCF) and the Spectral Flatness Measure (SFM) which provide complementary information about the spectral shape [15], [9]. These four descriptors are extracted by a frame analysis of 20ms windows length and 10ms hop size and concatenated, resulting in a 33 dimensional short term audio descriptor sequence (13 MFCC + 12 Chroma + 4 SCF + 4 SFM).

The temporal evolution of the obtained short term descriptors are then modeled by the following process: the amplitude spectrum of the temporal evolution of each component of the short term descriptors are computed. The obtained spectra are then passed through a filter bank and the log energy in each band are returned. The two types of global audio descriptors presented in this paper are extracted using two different filter banks. The first one, ID 11 in Table I, is composed of four

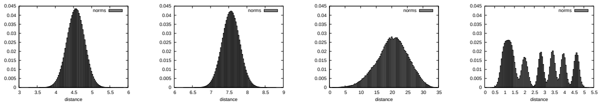
[1] European Web Archive (EWA) is an open archive that hosts several collections of public domain content crawled from publicly available resources.

ID	Descriptor type	Dimensionality	Distance function
Video Descriptors Datasets			
1	RGB global histogram	125	L_2
2	RGB global histogram	343	L_2
Audio Descriptors Datasets			
11		132	L_2
12		330	L_2
Synthetic Datasets			
101	random uniform	125	L_2
102	random uniform	343	L_2
Adopted Datasets			
201	ISOMAP face dataset	4096	L_2
202	Animal dataset	72	L_2

TABLE I
SUMMARY OF DATA THAT HAVE BEEN CONSIDERED FOR EVALUATION.

	Num. of Clips	Num. of Frames	Total Num. of Vectors
A	1	1,000	1,000
B	10	1,000	10,000
C	100	100	10,000
D	1000	10	10,000
E	10,000	1	10,000

TABLE II
SAMPLE SELECTION SUMMARY.



101 102 201 202

Fig. 1. Pairwise distance distributions of datasets ID 101, 102, 201 and 202.

rectangular filters centered in $[0, 1-2, 3-15, 20-43]Hz$. The second one, ID 12 in Table I, is composed of 10 rectangular filters, equally distributed in $[0, 43]Hz$. The two obtained global audio descriptors have 132 and 330 dimensions.

II. Datasets Recapitulation

Table I summarizes the datasets that were subject of our study. Besides the principal datasets of extracted audio and visual features, some auxiliary datasets have been also included in order to better understand and interpret the results of the study. In the last part of this section a sample selection method is discussed because the computational intensiveness of the criteria studied did not allowed direct application on the particular datasets as a whole.

A. Visual and Audio Datasets

As was mentioned earlier, the datasets were acquired processing the public domain content provided by EWA. In the case of visual datasets, about 2,000 hours of video were processed, computing the RGB global descriptor from one frame every two seconds. Thus, more than 10,000 videos were processed forming a dataset having about 3,500,000 feature vectors.

The audio descriptors were extracted from 10,000 musical audio files, totalizing 927 hours of signal. The temporal modeling was performed on a 3s window with a shift of 0.5s, producing 6,674,400 feature vectors.

B. Auxiliary Datasets

Synthetic floating type datasets with predefined number of dimensions and uniformly distributed were randomly generated. The values are ranging in an interval $[0, 1]$.

ISOMAP face dataset (ID 201) is a dataset of vectors representing synthetic faces used for evaluation of ISOMAP dimensionality reduction algorithm [13]. The dataset consists of 698 images (256 gray levels) of size 64 x 64 of the synthetic face where the rotation of the face and the lighting varies. The 4096-dimensional vectors represent linearized images which are compared using the L_2 metric. The last auxiliary dataset used is a dataset representing animals. It is clustered and represents kinds of animals where each is described by a 72-dimensional vector. It contains four clusters with 2,500 feature vectors, each one representing an instance of one animal. To which group the instance belongs is known. This dataset was used for classification methods evaluation for instance in [6].

C. Sample Selection Method

The criteria evaluation is often computationally intensive task that makes infeasible to use on the input the whole acquired dataset – millions of objects. This fact brings out the necessity to select sample from the whole dataset on which the evaluation will be done. In order to avoid biased results caused by improper or superficial sample selection, a set of sample datasets have been selected for each studied dataset.

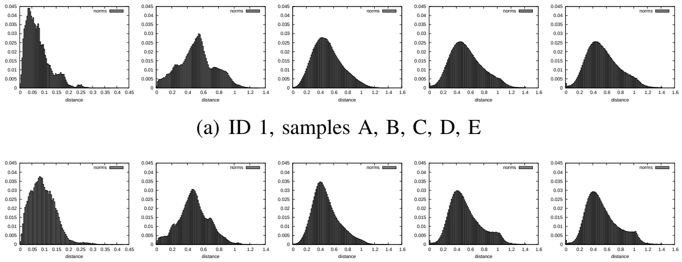
The samples are denoted using capitalized letters to distinguish from the IDs of the dataset and are summarized in Table II. The sample selection method takes into consideration the scale of redundancy in the sample. For instance, in sample type B, the 100 clips (either audio or video) were selected randomly from the whole dataset and from each clip, 100 frames were randomly selected, thus this sample has $100 \times 100 = 10,000$ feature vectors. For each dataset type (audio or video) the clip and frame selection was identical for both dimensionalities.

III. Pairwise Distance Distribution

The first criteria studied is the pairwise distance distribution. This criteria gives an insight into the organization of the distances among the feature vectors in the dataset. With comparison to the auxiliary datasets, whose structure is known, the overall structure of the data is discovered.

A. Auxiliary Datasets

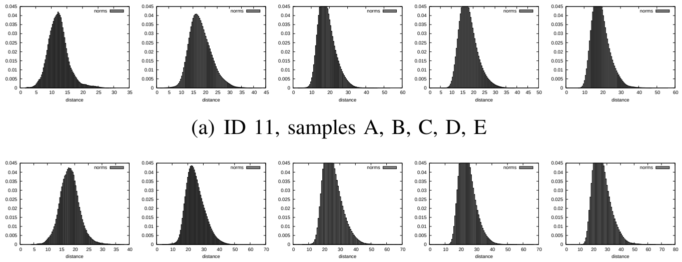
In Fig. 1, the pairwise distance distribution histogram of the animal dataset ID 202 shows six peaks in total. As mentioned in Section II-B, the dataset has four clusters with 2,500 feature vectors in each. There are six possible different combinations of distances among the cluster centers. Therefore, the first largest peak denotes both the distances within all clusters and the smallest distance between clusters that is 1.05. The other peaks denote the respective distance to the remaining data points in other clusters. The clusters are also well separated in the feature space since the distances from the cluster centers



(a) ID 1, samples A, B, C, D, E

(b) ID 2, samples A, B, C, D, E

Fig. 2. Pairwise distance distributions of the visual descriptor datasets.



(a) ID 11, samples A, B, C, D, E

(b) ID 12, samples A, B, C, D, E

Fig. 3. Pairwise distance distributions of the audio descriptor datasets.

are smaller than the distances among cluster centers – the peaks are significantly distinct.

The pairwise distance histograms of the other datasets ID 101, 102 and 201 in Fig. 1 do not exhibit behavior of such clustered dataset. They have only one peak and differ only by the indentation from zero, the height and the width of the peak – the variance. Simply, the narrower the peak is the more objects from the dataset have the more less the same distance from each other.

B. Visual Content

Fig. 2(a) and 2(b) depict the pairwise distance distributions for the visual descriptor datasets 1 and 2 respectively. The histograms of samples A to E are depicted from left to right. It can be seen that for the sample A, the redundancy is really high as it was expected, since the mutual distances among the objects in the dataset do not vary much in comparison with the other samples. In fact, the less redundant the sample is the larger is the range of the measured distances. This is desirable since it means that the histogram is not similar to that of the random datasets (ID 101 and 102, see Fig. 1). From the indexing point of view, high dimensional uniformly distributed random data is the hardest to index. It is due to the fact that vast majority of the objects have very similar mutual distance.

In general, no histogram of the visual dataset samples has the same shape as the histogram of the animal dataset (see Fig. 1, ID 202). In the case of sample B, small peaks can be seen at values 0.2 and 0.7, besides the large peak at value 0.45. Remember that sample B is formed by 10,000 frames taken from 10 videos – 1000 frames each. This might show that the intra-video frames are closer than the inter-video frames regarding the histogram of sample A. So, the videos might behave like clusters, yet the clusters are significantly overlapped in the space. An emerging peak at value 1 can be seen in histograms of sample E of datasets ID 1 and 2. This peak is formed by outliers and with the transition to greater dimensionality (ID 2) the interval of their distances to the rest of the dataset contracts.

The subject of the investigation is also the difference between the respective samples of the different dimensionality. From the mutual distances point of view we wanted to study whether with the growing dimensionality granularity of the descriptor was finer. Considering the respective pairs of samples, the shape slightly changes, yet both the mean and the variance of the histograms remain the same. The slight change is attributed to the greater amount of dimensions available in the case of the ID 2 dataset. Logical would be also larger mean in the case of the ID 2 dataset, yet the histogram prove that if there are some changes they are negligible. From these observations we can conclude that adding more than 200 dimensions to the ID 2 dataset did not bring any finer granularity and the information about the objects remained more less the same as in the case of the dataset ID 1.

C. Audio Content

Regarding the audio descriptor datasets, the results are depicted in Fig. 3. In this case, the shape of the histograms is more similar to the histograms of the random uniform datasets 101 and 102. Yet, still the variance of the histograms of the audio datasets is larger. On the other hand the mean, is also larger and according to [1] these are the main indicators of difficulties with the ability to index this data.

The enlarged dimensionality caused increase in the variance of the distances in the histogram, from which can be derived, that such implementation of the descriptor stores finer information about the original object. Yet, this must be verified by the intrinsic dimensionality estimation of these datasets.

IV. Intrinsic Dimensionality Estimation

To estimate the intrinsic dimensionality, two methods have been used: firstly, the linear PCA, secondly, an approach that is invariant to the non-linearity of the embedding.

A. Principal Component Analysis

PCA [14] is a statistical method that gives an insight into the internal structure of the data through its eigenvalue decomposition. It transforms the original vector data into lower dimensional data that respects its variance. In order to reduce the dimensionality, only the components of the eigenvalue decomposition, that significantly contribute to the data's energy (cumulative sum of the eigenvalues) are kept. Though, for each dataset minimal number of components needed to achieve the 95% of the energy of data was computed.

B. kNN Intrinsic Dimensionality Estimator

To estimate the intrinsic dimensionality, the estimator (kNN-IDE) described in [3] utilizes the notion of k-NN graph and its total length. The k-NN graph (kNNG) puts an edge between each point in the dataset ($\mathcal{X}$) and its k-nearest

ID	kNN-IDE					PCA				
	A	**B**	**C**	**D**	**E**	**A**	**B**	**C**	**D**	**E**
1	3.5	4.3	4.9	7.2	10.6	3	8	15	25	25
2	3.8	4.0	5.6	9.2	11.6	4	11	26	43	48
11	3.7	6.3	12.2	15.4	25	46	44	47	50	50
12	4.5	8.3	13.8	17.8	27	128	141	154	169	157
101	52.4					118				
102	92.8					319				
201	4.2					59				
202	12.1					11				

TABLE III
INTRINSIC DIMENSIONALITY ESTIMATIONS.

neighbors. Let $\mathcal{N}_{k,i}(\mathcal{X}_n)$ be the k-nearest neighbors of point $\mathbf{X}_i \in \mathcal{X}$, the total length of kNNG is defined as follows:

$$\hat{L}_\gamma(\mathcal{X}) = \sum_{i=1}^{n} \sum_{\mathbf{X} \in \mathcal{N}_{k,i}(\mathcal{X})} d^\gamma(\mathbf{X}, \mathbf{X}_i)$$

The authors in [3] found and proved the strong dependence of the length of the kNNG to the intrinsic dimensionality. Therefore they stated a simple estimator of the intrinsic dimensionality m:

$$\log \hat{L}_\gamma(\mathcal{Y}_n) = a \log n + b + \epsilon_n$$

where $\hat{L}_\gamma(\mathcal{Y}_n)$ is a total length of the kNNG of a uniform sample $\mathcal{Y}_n$, $a = (m - \gamma/m)$ and γ is power weighting constant, in our case $\gamma = 1$, b represents the entropy of the dataset and for the estimation of the intrinsic dimensionality is not necessary, ϵ_n is an error residual. a and b are approximated using several bootstrapping samples $\mathcal{Y}_n$ and using the method of moments and linear least squares. For our estimations we have used the same parameters as the authors. Each result of the estimation was rounded to the next greater integer and ten estimations were averaged to get the final estimation.

C. The Results

The results of the intrinsic dimensionality estimations are summarized in Table III. To interpret the results, the intrinsic dimensionality estimations of the auxiliary datasets needs to be explained at first. From [8] is known that the intrinsic dimensionality, in other words the degree of freedom, of the dataset ID 201 is three. Estimation using the kNN-IDE is 4.2, this discrepancy is attributed to different rounding method of the implementations. However, using PCA leads into significant overestimation of the intrinsic dimensionality. On the other hand, via PCA, the estimations of datasets 101 and 102 where the degree of freedom is very close to the actual dimensionality of the data is underestimated by the kNN-IDE.

By juxtaposing these observations for the studied datasets 1, 2, 11 and 12, two main conclusions can be derived. Firstly, the intrinsic dimensionality grows with the diminishing redundancy of the sample. Even though, for the least redundant sample E the intrinsic dimensionality does not reach the numbers of the datasets 101,102 and since it is the least redundant sample, this should represent the upper bound on the intrinsic dimensionality. Secondly, the estimations of the higher dimensionality dataset of the particular descriptor has similar estimation as the lower dimensionality dataset (kNN-IDE). Though, utilizing more dimensions for the extracted feature vector in this case does not mean having in the same extent finer granularity. In fact this could mean that with the same descriptiveness the acquired data would be worse suitable for indexing due to the fact it is embedded in much higher dimensional space.

V. CONCLUDING REMARKS AND FUTURE WORK

The main motivation behind this work was to get an insight into the internal structure of the high-dimensional multimedia data for further processing, for instance for searching. Especially for indexing, it is necessary to have the feature space as low dimensional as possible to reduce the curse of dimensionality. From this point of view the target dimensionality must be selected carefully because it seems that introducing more dimensions in the feature space necessarily not means finer granularity of stored information.

Even though the datasets represent euclidean vector spaces, the tools for this study, besides the PCA, have been selected with respect for other datasets and possible application to feature spaces where other distance functions might be employed.

As a future work we would like to verify the implications of this study by testing the indexability of the datasets presented here. The preliminary results of applying various indexing structures confirm the observations presented in this paper.

ACKNOWLEDGEMENTS

This work is done within the French Federation WISDOM (2007-2010) and is supported by the French project ANR MDCO DISCO (2007-2010).

REFERENCES

[1] E. Chávez and G. Navarro. A probabilistic spell for the curse of dimensionality. In *ALENEX '01: Revised Papers*, pages 147–160, London, UK, 2001. Springer-Verlag.

[2] B. Chazelle. Computational geometry: a retrospective. In *Proc. of ACM STOC*, pages 75–94, 1994.

[3] J. A. Costa and A. O. Hero. Geodesic entropic graphs for dimension and entropy estimation in manifold learning. *Signal Processing, IEEE Transactions on*, 52(8):2210–2221, 2004.

[4] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40(2), 2008.

[5] T. Deselaers, D. Keysers, and H. Ney. Features for image retrieval: an experimental comparison. *Inf. Retr.*, 11(2):77–107, 2008.

[6] J. H. Gennari, P. Langley, and D. Fisher. Models of incremental concept formation. *Artif. Intell.*, 40(1-3):11–61, 1989.

[7] Theo Gevers, Graham D. Finlayson, and Raimondo Schettini. Audio information retrieval: a bibliographical study. 2002.

[8] M. Hein and J.-Y. Audibert. Intrinsic dimensionality estimation of submanifolds in rd. In *ICML*, pages 289–296, 2005.

[9] J. Herre, E. Allamanche, and O. Hellmuth. Robust matching of audio signals using spectral flatness features. pages 127–130, 2001.

[10] B. S. Manjunath, P. Salembier, and T. Sikora. *Introduction to MPEG-7: Multimedia Content Description Interface*. Wiley & Sons, April 2002.

[11] G. Peeters, A. Laburthe, and X. Rodet. Toward automatic music audio summary generation from signal analysis. In *Proc. of ISMIR*, pages 94–100, Paris, France, 2002.

[12] M. J. Swain and D. H. Ballard. Color indexing. *International Journal of Computer Vision*, pages 11–32, November 1991.

[13] J. B. Tenenbaum, V. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000.

[14] M.E. Timmerman. Principal component analysis (2nd ed.). i. t. jolliffe. *Journal of the American Statistical Association*, 98:1082–1083, 2003.

[15] G. Wakefield. Mathematical representation of joint time-chroma distributions. In *Proc. of SPIE*, pages 637–645, Denver, Colorado, USA, 1999.