VISUAL FEATURE SPACE ANALYSIS FOR UNSUPERVISED EFFECTIVENESS ESTIMATION AND FEATURE ENGINEERING

Tobias Schreck   *Daniel Keim*

Databases and Visualization Group
University of Konstanz, Germany
{schreck,keim}@inf.uni-konstanz.de

Christian Panse

Functional Genomics Center
Uni | ETH Zurich, Switzerland
cp@fgcz.ethz.ch

ABSTRACT

The Feature Vector approach is one of the most popular schemes for managing multimedia data. For many data types such as audio, images, or 3D models, an abundance of different Feature Vector extractors are available. The *automatic* (unsupervised) identification of the best suited feature extractor for a given multimedia database is a difficult and largely unsolved problem. We here address the problem of *comparative unsupervised feature space analysis*. We propose two interactive approaches for the visual analysis of certain feature space characteristics contributing to estimated discrimination power provided in the respective feature spaces. We apply the approaches on a database of 3D objects represented in different feature spaces, and we experimentally show the methods to be useful (a) for unsupervised comparative estimation of discrimination power and (b) for visually analyzing important properties of the components (dimensions) of the respective feature spaces. The results of the analysis are useful for feature selection and engineering.

1. INTRODUCTION

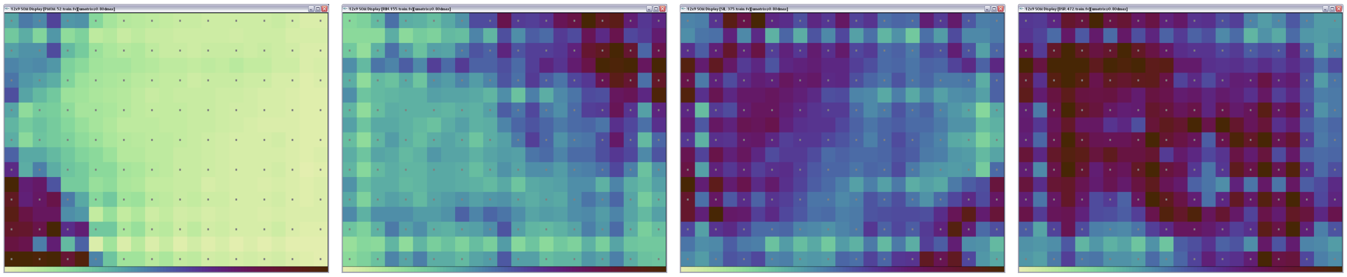
The Feature Vector (FV) approach [1] to managing multimedia data represents multimedia objects $o \in O$ given in an object space O by points $\vec{p_o} \in \mathbb{R}^d$ in a d-dimensional vector space. FV extractors fvx are functions $fvx : O \rightarrow \mathbb{R}^d$ mapping objects to vectors numerically describing object properties. Suitable extractors provide the generated FVs (a) are efficiently calculated and (b) allow to effectively capture object space similarity relationships by appropriate distance functions $d : (\vec{p_i}, \vec{p_j}) \rightarrow \mathbb{R}_0^+$ defined in FV space. The FV approach provides a simple, flexible means to implement important multimedia applications such as content-based retrieval and clustering. Also, the FV approach supports database indexing [2]. For many multimedia data types, description schemes other than FVs exit, e.g., relying on graph-based representations. Also, transformation-based matching schemes have been proposed for certain content. Yet, due to its simplicity and generality, the FV approach remains popular.

The *effectiveness* of a given FV extractor used to represent multimedia content is critical for any FV-based application. We understand the effectiveness of a FV extractor as the degree of how accurately distances d in FV space resemble similarity relationships in object space. For many multimedia data types an abundance of competing FV extractors are available. Yet often the identification of the most effective FV extractor for a given database is difficult. In this paper, we address this problem by proposing two visual tools for the comparative evaluation of FV spaces, and we demonstrate how the tools can support the selection and engineering of promising FV extractors from a pool of available FV extractors.

2. BACKGROUND

An abundance of FV extractors is evident for many important multimedia data types, e.g., in the image [3] and in the 3D model [4] domain. Effectiveness of FV extractors can be benchmarked if a suitable ground truth classification (supervised information) is available. Also, supervised FV engineering, e.g., by dimensionality reduction [5] or building appropriate combinations of FVs [6] is then possible. Practically, due to the large number of extractors available and the costs and even potential instability [7] associated with many benchmarks make supervised identification of the most effective FV extractors for a given application difficult. An alternative is to resort to unsupervised estimation of FV space effectiveness. To this end, a number of advanced statistical approaches have been proposed [8, 9]. These works are of rather theoretical nature and to the best of our knowledge have not been practically leveraged yet.

We here address the problem of unsupervised FV space analysis by means of characteristics obtained from compressed (clustered) FV space representations. As we are interested in visually supporting the analysis, we rely on the Kohonen (or Self-Organizing) Map algorithm [10] for FV space compression. It is a robust algorithm suited for visualization [11]. In [12] we applied Kohonen Maps in a multimedia retrieval system. Now, we leverage unsupervised information extracted from Kohonen Maps for FV space analysis and selection.



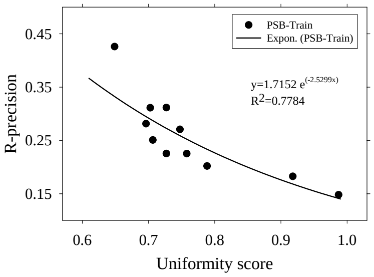
(a) PMOM (15%)

(b) RIN (23%)

(c) SIL (28%)

(d) DSR (43%)

Fig. 1. Visualization of the L_1 distances between adjacent cluster prototypes of Kohonen Maps generated for the PSB-Train database represented in four different feature spaces. Bright (dark) shades correspond to low (high) distances. The degree of uniformity of the respective distance distributions increases from left to right. This is in accordance with the increase of a supervised discrimination precision benchmark score (R-precision, given in brackets).



FV extractor	R-precision Score	Uniformity Score
DSR	42.61%	0.64872
DBF	31.16%	0.72692
VOX	31.13%	0.70256
SIL	28.15%	0.69615
CPX	27.08%	0.74744
3DDFT	25.08%	0.70641
GRAY	22.54%	0.72692
RIN	22.52%	0.75769
H3D	20.20%	0.78846
SD2	18.26%	0.91795
PMOM	14.82%	0.98718

Fig. 2. Regression analysis between uniformity score of Kohonen Map distance histograms (unsupervised information) and a supervised discrimination precision metric for eleven FV extractors. The expected correlation is verified, indicating viability of the analysis for automatic discrimination power estimation.

3. CLUSTER DISTANCE-BASED ANALYSIS

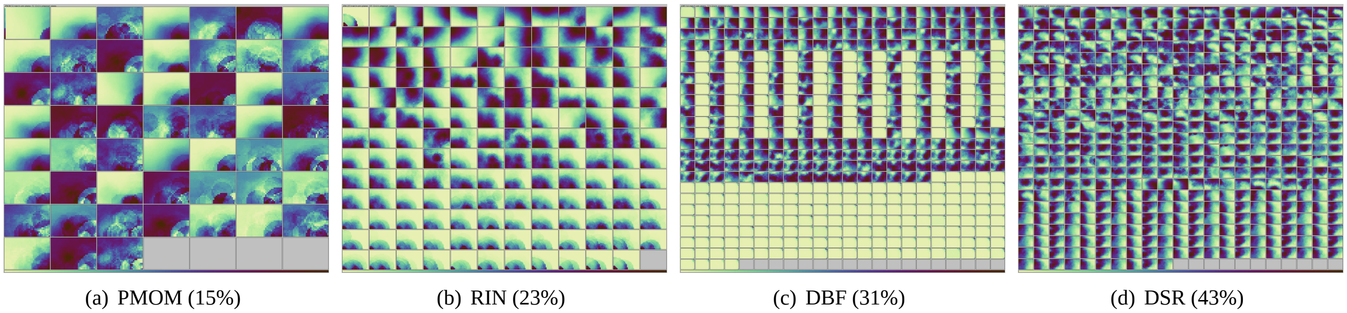
We propose an intuitive, simple, and practical method for unsupervised estimation of FV space discrimination power. We base our method on the following hypothesis:

Hypothesis 1 *Discrimination power provided in a given FV space can be estimated by the degree of uniformity of the distance histogram defined over inter-cluster distances in the respective FV space.*

An important assumption underlying Hypothesis 1 is that a FV space can be represented by a number of cluster prototypes as obtained by application of an automatic clustering algorithm, e.g., *k-Means* or the *Kohonen Map*. We then consider the distribution of distances between adjacent cluster prototypes. We expect the corresponding distance histograms to approximately resemble uniform distributions if the underlying FV spaces provide good discrimination power, as a-priori there is no rationality why any specific distance intervals should be preferred. While this has not necessarily to be the case for any possible combination of FV extractor and multimedia database, we expect uniform distance distributions to provide the best chances for meaningful discrimination in FV space. Conversely, we assume that for FV spaces providing only little discrimination power, cluster distances may be arbitrarily biased towards any subset of distance intervals.

We tested this hypothesis on a database of 3D models - the *Princeton Shape Benchmark* (PSB) Train partition [13] - described by a set of eleven competing FV extractors [4, 14]. We generated Kohonen Maps of dimensionality 12×9 for the database and each of the FV extractors. Figure 1 visualizes the distribution of L_1 distances between neighboring cluster prototypes on the Kohonen Maps for four different FV spaces. We note that we use L_1 as there are results that L_1 may be the most robust of the Minkowski distances for high-dimensional data [15]. In the respective images, brighter (darker) shades correspond to lower (higher) L_1 distances. From left to right, the degree of uniformity of the respective maps' distance distributions increases. While image (a) is dominated by low distances, image (d) consists of a rich mix of different distances. In terms of distance histograms, image (a) is skewed towards low distances, while image (d) approximately resembles a uniform inter-cluster distance distribution. Based on Hypothesis 1, we therefore expect the FV extractor underlying (d) to have best chances to provide good discrimination power, while we expect the converse for the FV extractor underlying (a). The two FV extractors of (b) and (c) should provide medium discrimination power as they show neither uniform nor extremely skewed distance distributions. Note that these assessments are based on unsupervised information automatically extracted from the respective FV spaces.

We verified these visually obtained effectiveness estimations by comparing them with benchmarked effectiveness scores obtained using the classification information accompanying the PSB database [13]. Specifically, we considered averaged *R-precision* scores [16] over the PSB in the different FV spaces. Briefly, R-precision is a measure for rating the quality of a retrieval algorithm based on a precision statistic over the



(a) PMOM (15%)

(b) RIN (23%)

(c) DBF (31%)

(d) DSR (43%)

Fig. 3. Component plane arrays for the PSB-Train database represented in four different feature spaces, sorted by benchmarked precision scores. The visualization allows unsupervised selection of prospective FV extractors, and can be used to identify highly correlated or indiscriminating components for removal from the FV. Note that the number of component planes differs among the arrays, as each FV extractor was equipped with a specific, method-dependent dimensionality setting.

answer lists returned when querying a labeled database for objects belonging to certain predefined object classes. Higher scores indicate better retrieval quality (better FV space effectiveness) regarding a given benchmark. The R-precision scores for each of the four FV extractors are included in Figure 1 and correlate positively with the degree of uniformity of the distance distributions.

We substantiate the above findings by a correlation analysis between R-precision scores and degree of uniformity of the Kohonen Map distance distributions given in the eleven FV spaces. For each FV space f, we calculate the *uniformity score* $us(h^f) = \sum_{i=1}^{b} |h_i^f - \frac{1}{b}|$ as the L_1-distance between its distance histogram h^f defined over b bins, and the uniform histogram of length b. The lower this score, the more uniform the resulting distance histogram is. Figure 2 gives the results of the exponential model regression analysis for the eleven FV extractors using $b = 8$ bin distance histograms. We verify the correlation between the supervised and the unsupervised FV space metric at squared correlation coefficient $R^2 = 0.78$. While this is not a perfect functional dependency, both metrics clearly correlate in the expected sense. We obtained similar results for different bin and Kohonen Map dimensionality settings. We conclude that the proposed analysis is a valid and practical option for addressing the unsupervised FV extractor selection problem.

4. COMPONENT-BASED ANALYSIS

Any meaningful distance function $d : (\vec{p_i}, \vec{p_j}) \rightarrow \mathbb{R}_0^+$ in vector space, such as the Minkowski or Quadratic Form distance functions, has to rely on the components (dimensions) in the FV space. So it is ultimately the sum of characteristics of the individual FV components that determines the FV effectiveness. We next state a second hypothesis, and propose a tool for visualizing certain component-based FV space characteristics supporting unsupervised discrimination power estimation and feature engineering.

Hypothesis 2 *Discrimination power provided in a given FV space can be estimated by the degree of heterogeneity among the components of the cluster prototype vectors representing the FV space.*

Similar to Hypothesis 1, the intuition behind Hypothesis 2 is that FV spaces exhibiting high heterogeneity of prototype vector components can be attributed better chances to provide meaningful discrimination power. The more biased the component values are towards certain component intervals, the less chances are expected for good discrimination power.

Based on these considerations, we propose interactive FV space evaluation by visualizing the component distributions of the cluster prototypes in FV space. Again, we rely on the Kohonen Map algorithm. A Kohonen *component plane* (CP) [11] visualizes the distribution of one selected FV dimension over the respective Kohonen Map. We can then visualize all component distributions in a FV space by simultaneously displaying the set of CPs in a component plane array (CPA).

Figure 3 shows CPAs of four different FV spaces (again, the PSB-Train database is used), ordered by increasing R-precision scores. Figure (a) contains the worst benchmarked FV extractor from our setting. Its CPA indicates that most components of the prototype vectors are biased towards certain value intervals, with substantial variance in component values only towards the bottom-right area of the CPs. We do not expect such characteristics to provide good chances for meaningful object discrimination. Conversely, image (d) corresponds to the most discriminative FV extractor according to the PSB benchmark. The respective CPA exhibits heterogeneous patterns for almost all components. We therefore are lead to expect good discrimination power.

Images (b) and (c) represent middle-ground situations regarding component heterogeneity. The extractor underlying image (b) exhibits significant variance among roughly the upper half of FV components. The lower half of components seem to be significantly correlated, as the respective CPs show similar patterns. Taking together these facts, we expect moderate discrimination power. A similar situation is present in image (c). About half of the components show significant variance, while the other half of the components represent roughly constant values which cannot meaningfully contribute to object discrimination. In this case, we note that the respective FV extractor was wrongly configured which lead to the observed outcome. Again, taking together both observations leads us to expect moderate discrimination power.

Besides discrimination power estimation, the CPA technique is also helpful in interactive FV engineering. The respective CPAs suggest that the highly correlated or approximately constant components can be aggregated or removed in the FVs underlying CPAs (b) and (c) in Figure 3. Doing so should lead to more compact FVs expected to retain the discrimination power provided by the original FVs.

We summarize that the CPA technique allows visual assessment of *variance*, *component-correlation*, and *noise / error* characteristics present among FV space components. In our experiments, we were able to verify these unsupervised, visually obtained assessments using supervised benchmarking results, indicating the usefulness of the CPA tool for FV selection and engineering. We note that numerically capturing the discussed CPA characteristics is difficult, and we leave the design of regression experiments similar to the one given in Section 3 for future work. We conclude that the CPA technique offers visual access to a wealth of useful FV space information.

5. CONCLUSIONS

We gave two hypotheses linking FV space characteristics obtained by unsupervised means with the discrimination power (effectiveness) to expect in the respective FV space. We gave experimental evidence supporting the hypotheses, and we demonstrated the applicability of two corresponding tools for visual FV space analysis. The tools are proposed to complement the (expensive) supervised benchmarking approach to FV space evaluation, and they are advocated for interactive FV selection and engineering tasks. The tools are specifically useful in cases where no appropriate benchmark is available.

Future work involves exploring additional unsupervised metrics for FV space discrimination power estimation. Besides the 3D FV domain considered in this work, we plan to apply the techniques in additional multimedia data domains.

6. ACKNOWLEDGMENTS

We thank Dietmar Saupe and Dejan Vranic for providing the 3D FV extractors and for valuable discussion. Thanks to Benjamin Bustos for helpful comments. This work was partially funded by the German Research Foundation (DFG) under Project No. KE 740/8-1.

7. REFERENCES

[1] C. Faloutsos, *Searching Multimedia Databases by Content*, Kluwer Academic Publishers, Norwell, MA, USA, 1996.

[2] C. Boehm, S. Berchtold, and D. Keim, “Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases,” *ACM Computing Surveys (CSUR)*, vol. 33, no. 3, pp. 322–373, 2001.

[3] R. Veltkamp and M. Tanase, “Content-based image retrieval systems: A survey,” Tech. Rep. UU-CS-2000-34, University Utrecht, 2000.

[4] B. Bustos, D. Keim, D. Saupe, T. Schreck, and D. Vranić, “Feature-based similarity search in 3D object databases,” *ACM Computing Surveys (CSUR)*, vol. 37, pp. 345–387, 2005.

[5] L. Yu and H. Liu, “Feature selection for high-dimensional data: A fast correlation-based filter solution,” in *Proceedings of The International Conference on Machine Leaning (ICML)*. 2003, pp. 856–863, AAAI Press.

[6] B. Bustos, D. Keim, D. Saupe, T. Schreck, and D. Vranić, “Using entropy impurity for improved 3D object similarity search,” in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME)*. 2004, pp. 1303–1306, IEEE.

[7] H. Mueller, S. Marchand-Maillet, and T. Pun, “The truth about corel - evaluation in image retrieval,” in *Proceedings of the International Conference on Image and Video Retrieval (CIVR)*. 2002, pp. 38–49, Springer.

[8] C. Aggarwal, “On the effects of dimensionality reduction on high dimensional similarity search,” in *Proc. ACM Symposium on Principles of database systems (PODS)*, 2001.

[9] A. Hinneburg, C. Aggarwal, and D. Keim, “What is the nearest neighbor in high dimensional spaces?,” in *Proceedings of the International Conference on Very Large Data Bases (VLDB)*, 2000, pp. 506–515.

[10] T. Kohonen, *Self-Organizing Maps*, Springer, Berlin, 3rd edition, 2001.

[11] J. Vesanto, “SOM-based data visualization methods,” *Intelligent Data Analysis*, vol. 3, no. 2, pp. 111–126, 1999.

[12] B. Bustos, D. Keim, C. Panse, and T. Schreck, “2D maps for visual analysis and retrieval in large multi-feature 3D model databases,” in *Proceedings of the IEEE Visualization Conference (VIS)*. 2004, IEEE Press, Poster paper.

[13] P. Shilane, P. Min, M. Kazhdan, and T. Funkhouser, “The princeton shape benchmark,” in *Proc. International Conference on Shape Modeling and Applications (SMI)*. 2004, pp. 167–178, IEEE CS Press.

[14] D. Vranić, *3D Model Retrieval*, Ph.D. thesis, University of Leipzig, Germany, 2004.

[15] C. Aggarwal, A. Hinneburg, and D. Keim, “On the surprising behavior of distance metrics in high dimensional spaces.,” in *Proceedings of the International Conference on Database Theory (ICDT)*, 2001, pp. 420–434.

[16] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison-Wesley, 1999.