Multimodal Analysis in Multimedia Using Symbolic Kernels

Hrishikesh B. Aradhye
SRI International, USA

Chitra Dorai
IBM T. J. Watson Research Center, USA

INTRODUCTION

The rapid adoption of broadband communications technology, coupled with ever-increasing capacity-to-price ratios for data storage, has made multimedia information increasingly more pervasive and accessible for consumers. As a result, the sheer volume of multimedia data available has exploded on the Internet in the past decade in the form of Web casts, broadcast programs, and streaming audio and video. However, indexing, search, and retrieval of this multimedia data is still dependent on manual, text-based tagging (e.g., in the form of a file name of a video clip). However, manual tagging of media content is often bedeviled by an inadequate choice of keywords, incomplete and inconsistent terms used, and the subjective biases of the annotator introduced in his or her descriptions of content adversely affecting accuracy in the search and retrieval phase. Moreover, manual annotation is extremely time-consuming, expensive, and unscalable in the face of ever-growing digital video collections. Therefore, as multimedia get richer in content, become more complex in format and resolution, and grow in volume, the urgency of developing automated content analysis tools for indexing and retrieval of multimedia becomes easily apparent.

Recent research towards content annotation, structuring, and search of digital media has led to a large collection of low-level feature extractors, such as face detectors and recognizers, videotext extractors, speech and speaker identifiers, people/vehicle trackers, and event locators. Such analyses are increasingly processing both visual and aural elements to result in large sets of multimodal features. For example, the results of these multimedia feature extractors can be

- *Real-valued*, such as shot motion magnitude, audio signal energy, trajectories of tracked entities, and scene tempo
- *Discrete* or *integer-valued*, such as the number of faces detected in a video frame and existence of a scene boundary (yes/no)
- *Ordinal*, such as shot rhythm, which exhibits partial neighborhood properties (e.g., metric, accelerated, decelerated)
- *Nominal*, such as identity of a recognized face in a frame and text recognized from a superimposed caption1

Multimedia metadata based on such a multimodal collection of features pose significant difficulties to subsequent tasks such as classification, clustering, visualization, and dimensionality reduction — all which traditionally deal with only continuous-valued data. Common data-mining algorithms employed for these tasks, such as Neural Networks and Principal Component Analysis (PCA), often assume a Euclidean distance metric, which is appropriate only for real-valued data. In the past, these algorithms could be applied to symbolic domains only after representing the symbolic labels as integers or real values or to a feature space transformation to map each symbolic feature as multiple binary features. These data transformations are artificial. Moreover, the original feature space may not reflect the continuity and neighborhood imposed by the integer/real representation.

This paper discusses mechanisms that extend tasks traditionally limited to continuous-valued feature spaces, such as (a) dimensionality reduction, (b) de-noising, (c) visualization, and (d) clustering, to multimodal multimedia domains with symbolic and continuous-valued features. To this end, we present four *kernel functions* based on well-known distance metrics that are applicable to each of the four feature types. These functions effectively define a linear or nonlinear dot product of real or symbolic feature vectors and therefore fit within the generic framework of kernel space machines. The framework of kernel functions and kernel space machines provides classification techniques that are less susceptible to overfitting when compared with several data-driven learning-based classifiers. We illustrate the usefulness of such symbolic kernels within the context of Kernel PCA and Support Vector Machines (SVMs), particularly in temporal clustering and tracking of videotext in multimedia. We show that such

analyses help capture information from symbolic feature spaces, visualize symbolic data, and aid tasks such as classification and clustering and therefore are eminently useful in multimodal analysis of multimedia.

BACKGROUND

Early approaches to multimedia content analysis dealt with multimodal feature data in two primary ways. Either a learning technique such as Neural Nets was used to find patterns in the multimodal data after mapping symbolic values into integers, or the multimodal features were segregated into different groups according to their modes of origin (e.g., into audio and video features), processed separately, and the results from the separate processes were merged by using some probabilistic mechanism or evidence combination method. The first set of methods implicitly assumed the Euclidean distance as an underlying metric between feature vectors. Although this may be appropriate for real-valued data, it imposes a neighborhood property on symbolic data that is artificial and is often inappropriate. The second set of methods essentially dealt with each category of multimodal data separately and fused the results. They were thus incapable of leading to novel patterns that can arise if the data were treated together as a whole. As audiovisual collections of today provide multimodal information, they need to be examined and interpreted together, not separately, to make sense of the composite message (Bradley, Fayyad, & Mangasarian, 1998).

Recent advances in machine-learning and data analysis techniques, however, have enabled more sophisticated means of data analyses. Several researchers have attempted to generalize the existing PCA-based framework. For instance, Tipping (1999) presented a probabilistic latent-variable framework for data visualization of binary and discrete data types. Collins and co-workers (Collins, Dasgupta, & Schapire, 2001) generalized the basic PCA framework, which inherently assumes Gaussian features and noise, to other members of the exponential family of functions. In addition to these research efforts, Kernel PCA (KPCA) has emerged as a new data representation and analysis method that extends the capabilities of the classical PCA — which is traditionally restricted to linear feature spaces — to feature spaces that may be nonlinearly correlated (Scholkopf, Smola, & Muller, 1999). In this method, the input vectors are implicitly projected on a high-dimensional space by using a nonlinear mapping. Standard PCA is then applied to this high-dimensional space. KPCA avoids explicit calculation of high-dimensional projections with the use of kernel functions, such as radial basis functions (RBF), high-degree polynomials, or the sigmoid function. KPCA has been successfully used to capture important information from large, nonlinear feature spaces into a smaller set of principal components (Scholkopf et al., 1999). Operations such as clustering or classification can then be carried out in this reduced dimensional space. Because noise is eliminated as projections on eigen-vectors with low eigen-values, the final reduced space of larger principal components contains less noise and yields better results with further data analysis tasks such as classification.

Although many conventional methods have been previously developed for extraction of principal components from nonlinearly correlated data, none allowed for generalization of the concepts to dimensionality reduction of symbolic spaces. The kernel-space representation of KPCA presents such an opportunity. However, since its inception, applications of KPCA have been primarily limited to domains with real-valued, nonlinearly correlated features despite the recent literature on defining kernels over several discrete objects such as sequences, trees, graphs, as well as many other types of objects. Moreover, recent techniques like the Fisher kernel approach by Jaakkola and Haussler (1999) can be used to systematically derive kernels from generative models, which have been demonstrated quite successfully in the rich symbolic feature domain of bioinformatics. Against the backdrop of these emerging collections of research, the work presented in this paper uses the ideas of Kernel PCA and symbolic kernel functions to investigate the yet unexplored problem of symbolic domain principal component extraction in the context of multimedia. The kernels used here are designed based on well-known distance metrics, namely Hamming distance, Cityblock distance, and the Edit distance metric, and have been previously used for string comparisons in several domains, including gene sequencing.

With these and other symbolic kernels, multimodal data from multimedia analysis containing real and symbolic values can be handled in a uniform fashion by using, say, an SVM classifier employing a kernel function that is a combination of Euclidean, Hamming, and Edit Distance kernels. Applications of the proposed kernel functions to temporal analysis of videotext data demonstrate the utility of this approach.

MAIN THRUST

Distance Kernels for Multimodal Data

Kernel-based classifiers such as SVMs and Neural Networks use linear, Radial Basis Function (RBF), or polynomial functions as kernels that first (implicitly) transform input data into a higher dimensional feature space and then process them in this space. Many of the common

kernels assume a Euclidean distance to compare feature vectors. Symbolic kernels, in contrast, have been less commonly used in the published literature. We use the following distance-based kernel functions in our analysis.

1. **Linear (Euclidean) Kernel Function for Real-valued Features:** This is the most commonly used kernel function for linear SVMs and other kernel-based algorithms. Let x and z be two feature vectors. Then the function

$$\mathrm{K}_l(x,z) = x^T z \tag{1}$$

defines a Euclidean distance-based kernel function. KPCA with linear kernel function reduces to the standard PCA. The linear kernel trivially follows Mercer's condition for kernel validity, that is, the matrix comprising pairwise kernel function values for any finite subset of feature vectors selected from the feature space is guaranteed to be positive semidefinite.

2. **Hamming Kernel Function for Nominal Features:** Let the number of features be N. Let x and z be two feature vectors, that is, N-dimensional symbolic vectors from a finite, symbolic feature space X_N. Then the function

$$\mathrm{K}_h(x,z) = N - \sum_{i=1}^{N} \delta(x_i, z_i) \tag{2}$$

where $\delta(x_i, z_i) = 0$ if $x_i = z_i$ and 1 otherwise, defines a Hamming distance–based kernel function and follows the Mercer's condition for kernel validity (Aradhye & Dorai, 2002). The equality, $x_i = z_i$, refers to symbol/label match.

3. **Cityblock Kernel Function for Discrete Features:** Using the preceding notation, the function

$$\mathrm{K}_c(x,z) = \sum_{i=1}^{N} M_i - |x_i - z_i| \tag{3}$$

where the ith feature is Mi-ary, defines a Cityblock distance–based kernel function and follows the Mercer's condition for kernel validity (Aradhye & Dorai, 2002).

4. **Edit Kernel Function for Stringlike Features:** We define an edit kernel between two strings as

$$\mathrm{K}_e(x,z) = \max(len(x), len(z)) - E(x,z) \tag{4}$$

where $E(x,z)$ is the edit distance between the two strings, defined conventionally as the minimum number of change, delete, and insert operations required to convert one string into another, and $len(x)$ is the length of string x. In theory, Edit distance does *not* obey Mercer validity, as has been recently proved by Cortes and coworkers (Cortes, Haffner, & Mohri, 2002, 2003). However, empirically, the Kernel matrices generated by the edit kernel are often positive definite, justifying the practical use of Edit distance–based kernels.

Hybrid Multimodal Kernel

Having defined these four basic kernel functions for different modalities of features, we are now in a position to define a multimodal kernel function that encompasses all types of common multimedia features. Let any given feature vector x be comprised of a real-valued feature set x_r, a nominal feature set x_h, a discrete-valued feature set x_c, and a string-style feature x_e, such that $x = [x_r \quad x_h \quad x_c \quad x_e]$. Then, because a linear combination of valid kernel functions is a valid kernel function, we define

$$\mathrm{K}_m(x,z) = \alpha \mathrm{K}_l(x_r, z_r) + \beta \mathrm{K}_h(x_h, z_h) + \gamma \mathrm{K}_c(x_c, z_c) + \delta \mathrm{K}_l(x_e, z_e) \tag{5}$$

where $\mathrm{K}_m(x,z)$ is our multimodal kernel and α, β, γ, and δ are constants. Such a hybrid kernel can now be seamlessly used to analyze multimodal feature vectors that have real and symbolic values, without imposing any artificial integer mapping of symbolic labels and further obtaining the benefits of analyzing disparate data together as one. The constants α, β, γ, and δ can be determined in practice either by a knowledge-based analysis of the relative importance of the different types of features or by empirical optimization.

Example Multimedia Application: Videotext Postprocessing

Video sequences contain a rich combination of images, sound, motion, and text. Videotext, which refers to superimposed text on images and video frames, serves as an important source of semantic information in video streams, besides speech, close caption, and visual content in video. Recognizing text superimposed on video frames yields important information such as the identity of the speaker, his/her location, topic under discussion, sports scores, product names, associated shopping data, and so forth, allowing for automated content description,

search, event monitoring, and video program categorization. Therefore, a videotext based Multimedia Description Scheme has recently been adopted into the MPEG-7 ISO/IEC standard to facilitate media content description (Dimitrova, Agnihotri, Dorai, & Bolle, 2000). To this end, videotext extraction and recognition is an important task of an automated video content analysis system. Figure 1 shows three illustrative frames containing videotext taken from MPEG videos of different genres.

However, unlike scanned paper documents, videotext is superimposed on often changing backgrounds comprising moving objects with a rich variety of color and texture. In addition, videotext is often of low resolution and suffers from compression artifacts. Due to these difficulties, existing OCR algorithms result in low accuracy when applied to the problem of videotext extraction and recognition. On the other hand, we observed that temporal videotext often persists on the screen over a span of time (approximately 30 seconds) to ensure readability, resulting in many available samples of the same body of text over multiple frames of video. This redundancy can be exploited to improve recognition accuracy, although erroneous text extraction and/or incorrect character recognition by a classifier may make the strings dissimilar from frame to frame. Often no single instance of the text may lead to perfect recognition, underscoring the need for intelligent postprocessing.

In the existing literature, unfortunately, temporal contiguity analysis of videotext is implemented by using ad-hoc thresholds and heuristics for the following reasons. First of all, due to missed or merged characters, the same string may be perceived to be of different lengths on different frames. We thus have feature vectors of varying lengths. Secondly, the exact duration of the persistence of videotext is unknown a priori. Two consecutive frames can have completely different strings. Thirdly, videotext can be in scrolling motion. Because multiple moving text blocks can be present in the same video frame, it is nontrivial to recognize which videotext objects from consecutive frames are instances of the same text.

In light of these difficulties, we present brief illustrative examples of dimensionality reduction and visualization of feature vectors comprising strings of recognized videotext. Experiments with our Edit distance kernel investigated the use of KPCA for analyzing the temporal contiguity of videotext using these feature vectors.

- **Videotext Clustering and Change Detection:** Figure 2 shows the first two components obtained by applying KPCA with the Edit distance kernel to a set of strings recognized from 20 consecutive frames. These frames contain instances of two distinct strings. Without any assumed knowledge, KPCA's use of the Edit distance kernel clearly shows two distinct clusters corresponding to these two strings.
- **Videotext Tracking and Outlier Detection:** Figure 3 shows the first three principal components obtained by applying KPCA with our Edit distance kernel to a set of strings recognized from 20 consecutive frames. These frames contain instances of videotext scrolling across the screen. In this three-dimensional plot, we can see a visual representation of the changing content of videotext as a trajectory in the principal component space, and locating outliers from the trajectory indicates the appearance of other strings in the video frames.

These results show that the symbolic kernels can assist significantly in automated agglomeration and tracking of recognized text as well as effective data visualization. In addition, multimodal feature vectors constructed from recognized text strings and frame motion estimates can now be analyzed jointly by using our hybrid kernel for media content characterization.

FUTURE TRENDS

One of the big hurdles facing media management systems is the semantic gap between the high-level meaning sought by user queries in search for media and the low-level features that we actually compute today for media indexing and description. Computational Media Aesthetics, a promising approach to bridging the gap and building

Figure 1. Illustrative frames with videotext

Figure 2. Videotext clustering

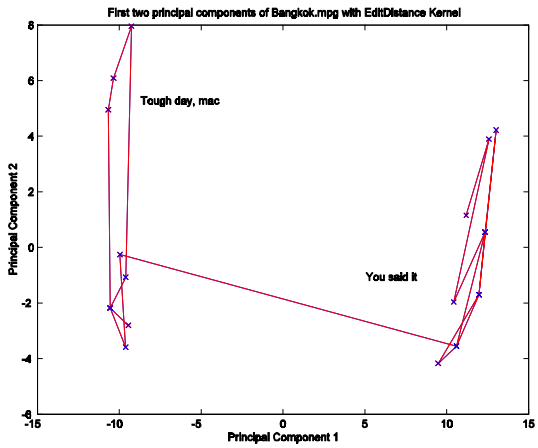
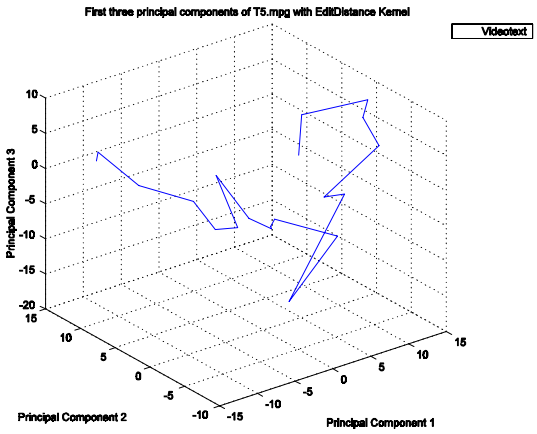


Figure 3. Videotext tracking and visualization



high-level semantic descriptions for media search and navigation services, is founded upon an understanding of media elements and their individual and joint roles in synthesizing meaning and manipulating perceptions, with a systematic study of media productions (Dorai & Venkatesh, 2001). The core trait of this approach is that in order to create effective tools for automatically understanding video, we need to be able to interpret the data with its maker's eye. In order to realize the potential of this approach, it becomes imperative that all sources of descriptive information, audio, video, text, and so forth need to be considered as a whole and analyzed together to derive inferences with certain level of integrity. With the ability to treat multimodal features as an integrated feature set to describe media content during classification and visualization, new higher level semantic mappings from low-level features can be achieved to describe media content. The symbolic kernels are promising an initial step in that direction to facilitate rigorous joint feature analysis in various media domains.

CONCLUSION

Traditional integer representation of symbolic multimedia feature data for classification and other data-mining tasks is artificial, as the symbolic space may not reflect the continuity and neighborhood relations as imposed by integer representations. In this paper, we use distance-based kernels in conjunction with kernel space methods such as KPCA to handle multimodal data, including symbolic features. These symbolic kernels, as shown in this paper, help apply traditionally numeric methods to symbolic spaces without any forced integer mapping for important tasks such as data visualization, principal component extraction, and clustering in multimedia and other domains.

REFERENCES

Aradhye, H., & Dorai, C. (2002). New kernels for analyzing multimodal data in multimedia using kernel machines. *Proceedings of the IEEE International Conference on Multimedia and Expo, Switzerland, 2* (pp. 37-40).

Bradley, P. S., Fayyad, U. M., & Mangasarian, O. (1998). *Data mining: Overview and optimization opportunities* (Tech. Rep. No. 98-01). Madison: University of Wisconsin, Computer Sciences Department.

Collins, M., Dasgupta, S., & Schapire, R. (2001). A generalization of principal component analysis to the exponential family. In T.G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in neural information processing systems 14* (pp. 617-624). Cambridge, MA: MIT Press.

Cortes, C., Haffner, P., & Mohri, M. (2002). Rational kernels. In S. Becker, S. Thrun, & K. Obermayer (Ed.), *Advances in neural information processing systems 15* (pp. 41-56). Cambridge, MA: MIT Press.

Cortes, C., Haffner, P., & Mohri, M. (2003). Positive definite rational kernels. *Proceedings of the 16th Annual Conference on Computational Learning Theory* (pp. 41-56), USA.

Dimitrova, N., Agnihotri, L., Dorai, C., & Bolle, R. (2000, October). MPEG-7 videotext descriptor for superimposed text in images and video. *Signal Processing: Image Communication, 16,* 137-155.

Dorai, C., & Venkatesh, S. (2001, October). Computational media aesthetics: Finding meaning beautiful. *IEEE Multimedia, 8*(4), 10-12.

Jaakkola, T. S., & Haussler, D. (1999). Exploiting generative models in discriminative classifiers. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in neural information processing systems 11* (pp. 487-493). Cambridge, MA: MIT Press.

Scholkopf, B., Smola, A., & Muller, K. R. (1999). Kernel principal component analysis. In B. Scholkopf, C. J. C. Burges, & A. J. Smola (Eds.), *Advances in kernel methods: SV learning* (pp. 327-352). Cambridge, MA: MIT Press.

Tipping, M. E. (1999). Probabilistic visualisation of high-dimensional binary data. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in neural information processing systems 11* (pp. 592-598). Cambridge, MA: MIT Press.

KEY TERMS

Dimensionality Reduction: The process of transformation of a large dimensional feature space into a space comprising a small number of (uncorrelated) components. Dimensionality reduction allows us to visualize, categorize, or simplify large datasets.

Kernel Function: A function that intrinsically defines the projection of two feature vectors (function arguments) onto a high-dimensional space and a dot product therein.

Mercer's Condition: A kernel function is said to obey Mercer's condition for kernel validity iff the kernel matrix comprising pairwise kernel evaluations over any given subset of the feature space is guaranteed to be positive semidefinite.

MPEG Compression: Video/audio compression standard established by Motion Picture Experts Group. MPEG compression algorithms use psychoacoustic modeling of audio and motion analysis as well as DCT of video data for efficient multimedia compression.

Multimodality of Feature Data: Feature data is said to be *multimodal* if the features can be characterized as a mixture of *real-valued, discrete, ordinal,* or *nominal values*.

Principal Component Analysis (PCA): One of the oldest modeling and dimensionality reduction techniques. PCA models observed feature data as a linear combination of a few uncorrelated, Gaussian principal components and additive Gaussian noise.

Videotext: Text graphically superimposed on video imagery, such as caption text, headline news, speaker identity, location, and so on.

ENDNOTE

[1] The term *symbolic* is loosely used in this paper to mean *discrete, ordinal,* and *nominal* features.