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