Dartmouth College

Dartmouth Digital Commons

Computer Science Technical Reports

Computer Science

7-10-1997

Multiple Media Stream Data Analysis: Theory and Applications (Extended version)

Charles B. Owen
Dartmouth College

Fillia Makedon
Dartmouth College

Follow this and additional works at: https://digitalcommons.dartmouth.edu/cs_tr

Part of the Computer Sciences Commons

Multiple Media Stream Data Analysis: Theory and Applications (Extended version)[1]

Dartmouth College Computer Science
Technical Report PCS-TR97-321

Charles B. Owen, Fillia Makedon

Dartmouth Experimental Visualization Laboratory
6211 Sudikoff Labs
Dartmouth College, Hanover, NH, 03755, USA
devlab@cs.dartmouth.edu

Abstract: This paper presents a new model for multiple media stream data analysis as well as descriptions of some applications of this model in development at Dartmouth College. This model formalizes the exploitation of correlations between multiple, potentially heterogeneous, media streams in support of numerous application areas. The goal of the technique is to determine temporal and spatial alignments which optimize a correlation function and indicate commonality and synchronization between media streams. It also provides a framework for comparison of media in unrelated domains. Applications such as text-to-speech alignment, functional magnetic resonance imaging, speaker localization, and degraded media realignment are described.

Keywords

MULTIMEDIA, INFORMATION RETRIEVAL, MULTIPLE STREAM ANALYSIS, SEMANTIC BANDWIDTH COMPRESSION, LIP SYNCHRONIZATION, SPEAKER LOCALIZATION

1 Introduction

This paper presents a new model for multiple media stream data analysis. Most common approaches to media data analysis can be described as "monomedia" approaches in that they focus on a single media stream. Examples include content based browsing of video by Arman et. al. (1994) and speech retrieval by Brown et. al. (1996). This work is important and complicated. However, many applications can benefit from analysis of multiple media streams. In such applications it is the relationship between streams which is important as opposed to the relationship between a query and a single stream.

---

[1] A shorter version of this paper appears in the Proceedings of Gesellschaft fur Klassifikation e.V., University of Potsdam, Potsdam, Germany, March 12-14, 1997

For the purpose of this research, *multiple media stream data analysis* is considered to be the derivation of temporal and spatial relationships between two or more media streams. One example is the correlation of lip motion to speech audio — the audio provides clues as to the motion of the lips. If the two streams can be *spatially* synchronized (moving lips located), several applications become possible. The audio can be used to predict the lip motion, allowing joint audio-video data compression, and the speaker can be physically located in the image sequence, providing locality cues for speaker recognition and robotic navigation.

An important application for multiple media stream data analysis is *cross-modal information retrieval*. Some media are far more difficult to query than others. Text queries are simple and highly accurate. The technology for very complicated information retrieval in text has been in place for many years. Speech, on the other hand, is very difficult to query. Locating words in audio streams is prone to all of the problems of open vocabulary word spotting and, therefore, requires relatively new technologies which are often proud to quote accuracy numbers in the 60% range (Brown, et. al. (1996)). If a textual transcription for recorded speech and its temporal alignment to the speech audio is available, a query into that text can be used to locate a desired speech segment. Computing the alignment between these two streams is an application of multiple media stream data analysis. The transcription provides a useful query mechanism for the audio: each query in the text is then used to locate the appropriate speech segments. The fundamental basis of cross-modal information retrieval is that one media is queried to produce a result in another. The underlying requirement is automatic synchronization.

Multiple media stream analysis is a relatively new area. Some early projects included alignment of newly recorded voice audio to degraded voice audio, a technique used for motion picture dubbing. This work is described along with many other unsolved application areas by Chen, et. al. (1995a). The ViewStation project included cross-modal retrieval of close-captioned video, which falls into the *script-tight* category described in this paper (Lindblad (1994)). Chen, et. al. (1995) illustrate joint audio-video coding based on audio prediction of lip motion in order to achieve higher compression levels.

2   Multiple media stream data analysis

This paper presents a general model for multiple media stream data analysis. This model provides a basis on which to build more practical models including a discrete formulation and a ranked results formulation. There are two primary goals of this analysis: spatial and temporal synchronization. Temporal synchronization derives the modification of the timing of one or more of the media streams in order to maximize the correlation. An example is the alignment of two similar audio streams using speed variation techniques so as to align the voices. In many applications one media will be considered to be on a reference

time frame and others temporally adjusted to achieve synchronization to that media. This is particularly the case when one media is in real time (measured in seconds) and the others are only causal in nature. Some applications get temporal synchronization for free and are only concerned with spatial synchronization.

Spatial synchronization is the second goal. This is a spatial translation of contents at a point in time in order to maximize correlation. This may consist of spatial warping or selection. *Warping* is the adjusting of parameters in order to rearrange a media element temporally or spatially. As an example, some motion analysis techniques attempt to compute the optimal spatial warping of images so as to cancel the motion from frame to frame. Selection can be considered to be warping wherein all of the unselected components of the media are omitted. Lip motion location by correlation with speech audio selects the particular moving components of the image sequence that represent the lips.

2.1   Continuous formulation

Equation 1 is the continuous formulation for multiple media stream data analysis.

$$\eta = \arg \max_{\tau_i \in T_i, \psi_i \in \Psi_i, i=1...N} \int_{-\infty}^{\infty} \rho(\psi_1(\mu_1 \circ \tau_1, t), ..., \psi_N(\mu_N \circ \tau_N, t)) dt \quad (1)$$

There are a large number of variables in this model. The model assumes N media streams are to be correlated, with functions $\mu_1, ..., \mu_N$ representing the media streams. The model assumes a media stream is a function of time. This is not a major restriction in practical discrete applications, as discussed in the next subsection. Each function τ_i is a *temporal synchronization* function and T_i is the set of possible temporal translation functions for stream i. These functions represent the possible time warpings of the media stream in order to achieve temporal synchronization. A selected temporal warping function τ_i must be a member of T_i.

ψ_i is a *domain translation* function and Ψ_i the set of all domain translation functions for stream i. A domain translation function translates complicated and highly redundant media data to a common and optimized comparison domain, where ρ, referred to as the *correlation* function, is used to produce the computed result η.

Domain translation is an important element of the model. It is difficult to construct appropriate correlation functions with disparate media parameters; as an example, what does it mean to compare audio, a one-dimensional function, to video, a two-dimensional function? However, comparison of estimated lip movement in audio to motion vectors in video is much more easily realized. Even when the media are the same, they may be too complicated to correlate

3

directly. For this reason, it is specified in this model that all media be translated to a common domain for comparison by the correlation function ρ.

In many cases ρ is some derivative of simple statistical correlation. Applications of multiple media stream correlation undertaken at the DEVLAB have favored complicated domain translation and simple correlation.

It is important to note that the result of this computation is not the maximization of the integral, but, rather, the *functions which achieve this maximization*. This model specifies a structure for defining and selecting functions which maximize the correlation. The resulting functions indicate necessary temporal and spatial correlations. The set of temporal translation function results $\tau_{i,i=1,\ldots,N}$ indicate the appropriate temporal synchronization for maximum correlation. The set of domain translation function results indicate the appropriate spatial synchronization for maximum correlation.

There are several unique characteristics of this model which should be noted. The parameters of the spatial synchronization function are time and the composition of the media function and the temporal synchronization function. This allows flexibility in the construction of the spatial synchronization function because it can compute synchronization based on a *neighborhood* of the time point. Many technologies, including simple filtering, require access to a neighborhood like this.

2.2 Discrete formulation

Equation 2 is the discrete formulation for multiple media stream data analysis. This is derived from the continuous formulation with the constraint that the media functions $\mu_{i,i=1,\ldots,N}$ are discrete functions. While the continuous formulation of this problem is primarily of theoretical interest, the discrete formulation can be practically implemented in discrete computer systems.

$$\eta = \arg \max_{\tau_i \in T_i, \psi_i \in \Psi_i, i=1\ldots N} \sum_{-\infty}^{\infty} \rho(\psi_1(\mu_1 \circ \tau_1, t), \ldots, \psi_N(\mu_N \circ \tau_N, t)) \qquad (2)$$

This model is compatible with non-temporal media provided the media is causal. As an example, text is not temporal, but does have a definite ordering. This ordering can be considered to be equivalent to a temporal ordering in the model. The temporal synchronization functions are not restricted to those requiring inverses, and can be used to reorder content if necessary (as in media presentations where content is repeated).

The most extreme case of non-temporal media is hypertext, wherein content can be represented as a graph with media nodes and navigation edges. This case is not supported by this model and would require modifications to account for the navigation edges.

4

2.3 Additional formulations

The multiple media stream data analysis model is very general and lends itself well to additional derivations. One such derivation is the *discrete ranked results* formulation. In this formulation the result is an ordered vector $< \eta_1, ..., \eta_M >$ of size M such that η_1 represents the best result and η_j represents the jth best result. In information retrieval it is common that results are considered to be probabilistic with decreasing probability as the matching score decreases. Increased probability of finding the correct match is achieved by increasing the match set size.

2.4 Computational approaches

These models provide a basic modularization and theoretical description of the multiple media stream data analysis and correlation problem. They do not specify the functions to use for any particular application. This paper describes how appropriate functions have been selected for several applications.

In general, the function sets in the model are implemented using *parameterization*. Rather than selecting from an often infinite set of functions, an impossible task in real systems, each function set is modeled using a single function with adjustable parameters. These parameters may be discrete or continuous. A common example parameterization is the set of temporal warpings of a media stream. Each discrete time point following the temporal warping is the result of function τ. This function can be implemented as a vector of temporal offsets (temporal flow) or as curve parameters (such as a polynomial). The solution reduces to iterative selection of parameters which maximize the function.

3 Applications

Several applications of multiple media stream correlation have been undertaken at the DEVLAB, both in order to further the theory of the model and as significant projects in their own right. This section briefly describes a few of these projects and how the model applies to them.

3.1 Text-to-speech synchronization

A major project at the DEVLAB is *Speak Alexandria*. The goal of Speak Alexandria is information retrieval in speech-based media. Speech-based media is any media which is predominately human speech. This is an interesting media in that it is very common and can be queried using simple text queries. Speech-based media can be divided into three types: script-less, script-light, and script-tight. Script-less media is media for which no transcription exists (live recordings, etc.). Script-tight media is media for which a tightly synchronized

transcription exists (such as close-captioned video). Script-less media requires voice recognition technologies for query processing and is not the subject of this paper. Script-tight content is a simple application of cross-modal information retrieval.

Script-light media is speech-based audio content for which an unaligned transcription exists. This is a surprisingly large category of content including broadcasting, dramatic performances, and court proceedings. This content can be queried efficiently provided the synchronization between the text and the speech can be computed.

There is significant advantage to treating script-light material as cross-modal information retrieval rather than applying the same techniques used for script-less content. The script represents the actual content and can drive voice recognition tools with the words that exist in the content rather than attempting to automatically recognize the text and produce a new transcription. Much higher accuracy is possible using this approach.

This application has two media streams. μ_1 represents the text and μ_2 represents the audio. In this application the temporal synchronization function τ_2 is the identity function, i.e. the text will be aligned to the speech. The alignment is easily invertible to provide speech to transcription retrieval if necessary (given a location in the audio, provide the written transcription at that point).

Ψ_1 is implemented by converting the text to a biphone graph. Phonemes are units of pronunciation in language. Biphones represent the transition between phonemes (as well as a context independent "middle" of the phoneme). This is a common technology used in voice recognition (See Rabiner (1993)). A directed graph is constructed of biphone translations of the words in the text wherein edges represent the possible biphone transitions. In the current implementation an optional .pau (pause) biphone is placed between words. Since many words have multiple pronunciations ("the" for example), the graph has multiple paths. In addition, *bypass edges* are included to model errors in the transcription (skipped or substituted content). This constructed graph is called a *transcription graph* and is represented in the multiple media stream correlation model as Ψ_1. It is a goal of the computation to select the correct path ψ_1 through this graph for the speech audio. The duration of individual biphones is not known, so explicit edges are added to each node which loop back to the node itself.

Ψ_2 is implemented using speech recognition tools which convert audio into a sequence of *biphone probabilities*. This process is beyond the scope of this paper, but is described in detail in Rabiner (1993) and Hermansky (1991). In summary, audio is blocked into finite frames (of 10 millisecond duration) and the probability of any given biphone at that point is time is computed. In the current implementation 536 biphone probabilities are computed for each audio frame. These vectors are treated as a lattice where each $\psi_2 \in \Psi_2$ is a left-to-right path through the lattice. The optimal ψ_2 is a path which is valid in the transcription graph and maximizes the total path probability. Clearly, ψ_1 and

6

ψ_2 must correspond.

The temporal synchronization function τ_1 is computed from the path ψ_1 through the transcription graph. In effect, these parameters are computed simultaneously. Computation of this path can be performed using the Viterbi algorithm, which is described in Rabiner (1993). In order to prevent numeric underflow, all probabilities are computed as logarithms. In this implementation, $\rho(\alpha, \beta) = \alpha + \beta$. Several enhancements of the Viterbi algorithm have been included in this application including modification of probabilities in the transcription graph over time to force interpolation and path pruning to decrease computational complexity.

3.2   Functional magnetic resonance imaging

Functional magnetic resonance imaging (fMRI) is a new application of MRI technology which captures rapid sequences of images of internal body function (Bandettini (1997)). The DEVLAB has been working with the Dartmouth Medical School on fMRI imaging of the human brain. Using equipment now available, a complete brain volume image can be collected once every two seconds. The system can be tuned to discriminate between oxyhemoglobin and deoxyhemoglobin, a differentiation that has been shown to indicate brain activity.

In this application, a subject is placed in the scanner and a three minute sequence of images are captured (192 repetitions). The subject is provided with a multimedia presentation, either auditory or audio/visual using headphones and special eye goggles. The goal of this research is to localize activations in the brain induced by the multimedia presentation.

In applying multiple media stream correlation to this application, μ_1 is the multimedia *stimulus*. μ_2 is the fMRI data. ψ_1 is supplied and translates the multimedia presentation into a stimulus sequence. The goal is to compute the spatial localization ψ_2 and the hemodynamic response characteristic of the brain τ_1. The correlation function is simple statistical correlation. This work has enormous application in medical diagnostics. Related work is on the verge of replacing the existing invasive Wada test, an anesthetizing of one brain hemisphere used to localize the speech centers.

4   Other Applications

There are many other potential applications for multiple media stream correlation. Speaker localization is the identification of a speaker in an image sequence. This task can be performed by correlating speech audio to motion vectors in the image sequence in order to locate moving lips. Preliminary work in this area is very promising. This technique can support speaker retrieval in sequences as well as navigation in areas such as robotics.

There are many applications for media alignment. An early multiple media alignment system called WordFit described in Chen (1996a) has been in use in motion picture production for many years. WordFit aligns speech audio temporally to other speech audio. It is used to align a newly "dubbed" audio track to the original, where the original is deemed unsuitable due to poor audio quality or bad acting. WordFit is based on alignment of speech to speech. An open problem is the alignment of speech to silent video as well as the alignment of video to speech. The first method is useful for restoration of films where the audio track has been lost. The second is useful for foreign language dubbing.

5   Summary

This paper describes multiple media stream correlation, a general model for the spatial and temporal alignment of multiple media streams. This model provides a framework for research on retrieval and analysis algorithms which work on more than one media stream, gaining useful and important information from the relationships between these streams. Several applications areas are described which illustrate the wide range of uses for this technology. It is a technique with an open future, and the basis for much DEVLAB research at this time.

References

ARMAN, F., DEPOMMIER, A. and HSU, A., and CHIU, M.Y. (1994): Content-based browsing of video sequences. Proc. of ACM Multimedia'94, 97-103, San Francisco, CA.

BANDETTINI, P. A. and WONG, E. C. (1997): Echo Planar Imaging, chapter in *Echo-Planer Magnetic Resonance Imaging of Human Brain Activation*. Springer-Verlag.

BROWN, M. G., FOOTE, J. T., JONES, G. J. F., SPARCK JONES, K. and YOUNG, S. J. (1996): Open-vocabulary speech indexing for voice and video mail retrieval. *Proc. of Multimedia'96*, 307-316, Boston, MA.

BLOOM, P. J. and MARSHALL, G. D. (1984): A Digital Signal Processing System for Automatic Dialogue Post-Synchronization. *SMPTE Journal, Volume 93, Number 6, June, 1984, 566–569.*

CHEN, T. and RAO, R. (1995): Audio-Visual Interaction in Multimedia: From Lip-Syncronization to Joint Audio-Video Coding. *IEEE Circuits and Devices*, 11(6), November, 1995, 21–26.

CHEN, T. and RAO, R. (1995a): Cross-Modal Prediction in Audio-Visual Communication. *Proc. ICASSP'96*, Atlanta, GA, May, 1996.

HERMANSKY, H., MORGAN, N., BAYYA, A., and KOHN, P. (1991): *RASTA-PLP speech analysis*. Technical Report TR-91-069, International Computer Science Institute, Berkeley, CA.

LINDBLAD, C. J. (1994): A programming system for the dynamic manipulation of temporally sensitive data. MIT/LCS/TR-637, Massachusetts Institute of Technology.

RABINER, L. and JUANG, B.-H. (1993): *Fundamentals of Speech Recognition*, Signal Processing Series, PTR Prentice Hall, Englewood Cliffs, NJ.