Audio-Visual Group Recognition Using Diffusion Maps

Yosi Keller, Ronald R. Coifman, Stéphane Lafon, and Steven W. Zucker, *Fellow, IEEE*

***Abstract*—Data fusion is a natural and common approach to recovering the state of physical systems. But the dissimilar appearance of different sensors remains a fundamental obstacle. We propose a unified embedding scheme for multisensory data, based on the spectral diffusion framework, which addresses this issue. Our scheme is purely data-driven and assumes no *a priori* statistical or deterministic models of the data sources. To extract the underlying structure, we first embed separately each input channel; the resultant structures are then combined in diffusion coordinates. In particular, as different sensors sample similar phenomena with different sampling densities, we apply the density invariant Laplace–Beltrami embedding. This is a fundamental issue in multisensor acquisition and processing, overlooked in prior approaches. We extend previous work on *group recognition* and suggest a novel approach to the selection of diffusion coordinates. To verify our approach, we demonstrate performance improvements in audio/visual speech recognition.**

***Index Terms*—Dimensionality reduction, multisensor, sensor fusion, speech recognition, Laplacian eigenmaps.**

I. Introduction

IT is natural to look at a speaker with a thick accent in a noisy environment. Moreover, it is disturbing to watch an audio-video sequence in which the voice signal is delayed. These two familiar observations illustrate the advantage of sensor fusion as well as the challenge of achieving it in an uncontrolled environment. That it is useful is clear from neurobiology (see, e.g., [1] and [2]), demonstrating that it is feasible for audio-visual fusion is our goal. To achieve this, we extend existing sensor fusion algorithms by proposing a new approach based on mathematical embedding techniques. Performance improvements are demonstrated on the audio-visual task. The key is a proper abstraction of the pure signals to isolate relevant structures from individual streams before combining them to articulate their common structure and proceeding with classification.

While single-sensor systems have been successfully used for certain well-defined measurement tasks (e.g., blood pressure sensors in biomedicine), there are many applications in which a single sensor is almost never sufficient. In medical imaging—for instance, X-ray, computed tomography (CT), and magnetic resonance imaging (MRI)—acquire different physical and metabolic properties; successful diagnosis can depend critically on information inferred from their combination. As another example, the high spatial resolution and low signal-to-noise ratio of optical sensors in remote sensing are often complemented by radar-based synthetic aperture radar sensors in complex weather situations to remove atmospheric effects. The multisensor approach allows the resolution of ambiguities and the reduction of uncertainties, and may even fill in missing information.

The proposed data fusion scheme is as follows. Given k *synchronized* sources $s_1, s_2, \ldots, s_k$ of data, our goal is to reduce the dimensionality of the collected data by computing a low-dimensional composite representation. The term *synchronized* implies that, for a given epoch, the outputs of all of the sources are related to the same state of the observed system, as is the case for the audio and video streams of a single speaker.

To introduce and motivate our approach, consider the multisensor data set depicted in Fig. 1. The two synchronized image sets correspond to a person rotating his head from left to right. We consider each image as a signal in a high-dimensional space (of $O(10^4)$). (In a real application, the image set might contain X-ray and MRI images.) Despite the dissimilar appearance, both sets are the manifestations of the same phenomenon (a rotating head) and the same latent variable. The variation between the images is essentially one dimensional, as the rotation can be parameterized by a single rotation angle. We are not interested in the spatial domain pixelwise variation within a single image but in the variations between the different images.

By reducing the dimensionality of each input channel, we aim to recover a common parametrization corresponding to the rotation angle. Thus, the multisensor embedding might improve the recognition of configurations related to the head position (rotation angle).

The dissimilar appearance also implies that the density of both sets will differ. More particularly, consider a kernel density estimation scheme [3] applied to each set. Since the distances between images in each set will differ, so will the density. Spectral embeddings (parameterizations) are known to depend on the density of the embedded data set [4]. In order to find a common ground in terms of density, the natural choice is to normalize the density of each set and embed the sets under *uniform density*. This notion is formalized by the Laplace–Beltrami embedding and presented in Section III-A.

In this paper, we provide two main contributions. In our core contribution, we propose to utilize diffusion maps [4], [5] to de-

Manuscript received September 17, 2008; accepted July 17, 2009. First published August 21, 2009; current version published December 16, 2009. The associate editor coordinating review of this manuscript and approving it for publication was Prof. P. K. Varshney. This work was supported by AFOSR, ARO, and NGA.

Y. Keller is with the School of Engineering, Bar Ilan University, Israel (e-mail: yosi.keller@gmail.com).

R. R. Coifman is with the Department of Mathematics, Yale University, New Haven, CT 06520 USA (e-mail: coifman@math.yale.edu).

S. Lafon is with Google Inc., Mountain View, CA 94043 USA (e-mail: stephane.lafon@gmail.com).

S. W. Zucker is with the Department of Computer Science, Yale University, New Haven, CT 06520 USA (e-mail: steven.zucker@yale.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSP.2009.2030861

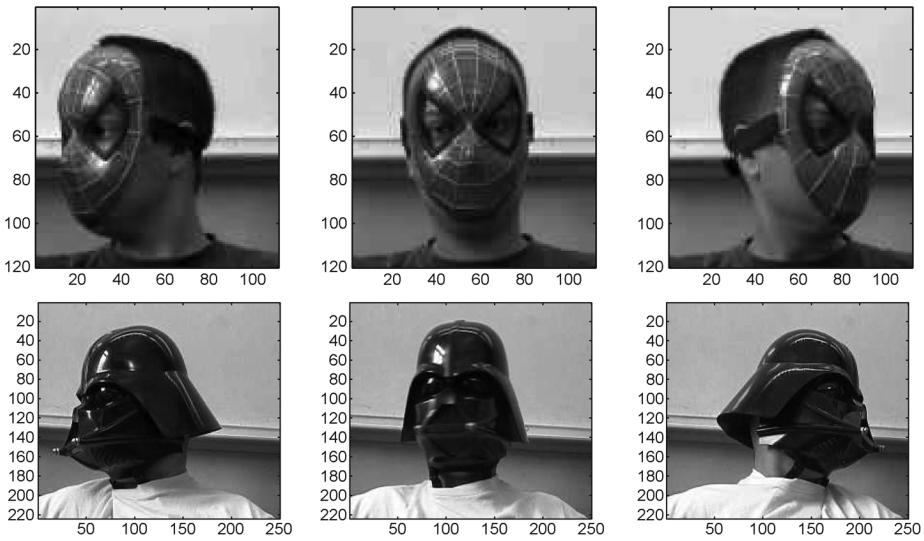


Fig. 1. Multisensor data sets related to a common manifold. For each rotation angle of the head we are given two images, one from each row. We seek the common manifold, which encodes the head position (rotation angle).

rive a unified representation of heterogenous data sources, corresponding to the same underlying phenomenon and manifold. Since our data sources are physically different, as with audio and video, they cannot be combined in their original form. However, their low-dimensional embeddings are in a common form that, most importantly, can be composed. We apply these approaches to audio-visual speech recognition and show that the use of heterogeneous data sources improves accuracy when compared to any single data source. The specific importance of density-invariant embeddings [6] and their use in pattern recognition applications is underscored. This is the new application contribution in the paper.

Secondly, based on the proposed fusion scheme, we present a multisensor supervised pattern recognition framework for "group objects." We apply the Hausdorff distance[1] to learn and recognize "group objects" and derive a *diffusion coordinates* selection scheme. Namely, given the diffusion coordinates (spectral embedding) of a heterogeneous or homogeneous data source, we suggest an algorithm for choosing a subset of the coordinates that maximizes the Hausdorff classification margin and improves the recognition rate.

This paper is organized as follows: a survey of prior art in multisensor data analysis is given in Section II. We summarize density invariant embeddings and spectral extensions in Section III, while two complementary approaches for the selection of the kernel bandwidth are presented in Section III-C. Our extension to *group recognition* is described in Section V, and a new approach to diffusion coordinates selection is given in Section VI. We exemplify and verify our approach experimentally by applying it to audio-visual speech recognition in Section VII. Summary and concluding remarks are given in Section VIII.

II. Background

Multisensor data analysis and fusion is discussed in multiple disciplines such as signal processing, biology, and image processing, to name a few. As a result, many different techniques have been proposed, which can be categorized into addressing two fundamental versions of the problem, the first being to recover an underlying process and the second for task-oriented fusion schemes. As we show, our technique builds upon an approach in the first class but is applicable to both. We stress, however, that only by defining proper classification methodologies can it be made to work at acceptable rates.

The first class of problem involves data-driven schemes designed to recover an underlying process that is manifested by the multisensor data. Multisensor registration of images is perhaps a paradigm example. In multimodal medical imaging (MRI, CT), for example, the different image-acquisition processes give rise to individual pixel (voxel) relationships that are complex and difficult to model. Hence, multisensor image alignment is typically based on deriving canonical image representations that are invariant to the dissimilarities between the sensors yet capture relevant image structure. The underlying structures are geometric primitives, such as feature points, contours, and corners [7]–[9]. In our technique, of course, we do not have to define these primitives *a priori*, but they should emerge implicitly from the embedding.

In such approaches, it is common to model the underlying structure as a random variable or process. Thus, the different sensors are assumed to be different manifestations of the same random variable/process. For instance, in Viola's work on multisensor image registration [10], multisensor images are modeled as random variables. The motion parameters are estimated by maximizing the mutual information (MI) between the different image modalities, with respect to the motion parameters. The nonparametric probabilities were computed by Parzen windows. Ozertem and Erdogmus [11] used MI to compute pairwise affinities between sensors and thus derive an importance-weighting scheme for each sensor. The scheme is shown to improve the performance of a particle filter–based motion tracker.

Another class of approaches involve schemes based on dimensionality reduction. These are of particular interest to us,

[1]http://en.wikipedia.org/wiki/Hausdorff_distance/.

since our approach is in this class. For these approaches, the notion of the statistical model is replaced by the *data manifold*. Low-dimensional manifolds can then be parameterized and unified. A general-purpose approach to high-dimensional data embedding and alignment was presented by Ham *et al.* [12]. Given a set of *a priori* corresponding points in the different input channels, a constrained formulation of the graph Laplacian embedding is derived. First, they add a term fixing the embedding coordinates of certain samples to predefined values. Both sets are then embedded separately, where certain samples in each set are mapped to the same embedding coordinates. Secondly, they describe a dual embedding scheme, where the constrained embeddings of both sets are computed simultaneously, and the embeddings of certain anchor points in both datasets are constrained to be identical.

Canonical correlation analysis (CCA) is a statistical approach that combines linear dimensionality reduction and fusion. Given two input sources x and y, CCA computes linear projections of x and y such that the projections are maximally correlated. Kidron *et al.* [13] applied CCA to multisensor event detection, thus deriving a common parametrization of the audio and visual signals. The authors show that the CCA is ineffective in analyzing high-dimensional data, as the estimation of the corresponding covariance matrix becomes statistically ill-posed. The issue was resolved by applying a sparsity prior, as the common manifold related to audio/visual events was sparse. CCA was applied by Sargin *et al.* [14] to open-set speaker identification by integrating speech and lip texture features. The authors also propose a method for synchronizing the speech and lip features using CCA. Although our foundation is different, we shall show how CCA can be utilized more effectively.

Spectral embedding was applied to multisensor *synchronization* by Lafon *et al.* [15]. They showed that unsynchronized high-dimensional data (images) of the same phenomenon (rotating heads) acquired by different sensors had a common low-dimensional manifold, and that this common manifold could be recovered by the Laplace–Beltrami embedding. (This motivated our opening illustration.) The low-dimensional embeddings of the different sensors were then coregistered to synchronize the rotations of the heads. In this paper, we assume the sensors to be synchronized à la Lafon but further utilize sensor fusion to improve recognition. Our classifier design built onto diffusion embeddings thus extends the Lafon paper significantly.

Another class of algorithms is based on multiple kernel learning (MKL) and was introduced by Bach *et al.* [16], [17]. These approaches can be used to derive multisensor *SVM classifiers*. In contrast, CCA and the proposed scheme provide a *unified representation* of the multisensor data that can then be used by *any* data analysis scheme, not necessarily a binary classifier. Similar to our approach, a different kernel is applied to each data source, and the fusion of the different sources is achieved by summing the kernel matrices corresponding to each data source. Lanckriet *et al.* [18] extended the MKL by deriving an SVM classifier using a kernel matrix that was the weighted sum of the kernel matrices corresponding to the different data sources $K = \sum_i w_i K_i$. The weights $\{w_i\}$ were optimized as part of the SDP optimization used to derive the SVM classifier's parameters. These approaches were applied to biological data [18], [19]. Gene functional classifications were inferred by Pavlidis *et al.* [19] by applying MKL-SVM classifiers to a heterogeneous data set consisting of phylogenetic profiles and DNA microarray expression measurements. Lanckriet *et al.* [18] analyzed Yeast functionals that were clustered by a weighted MKL-SVM classifier, trained on a heterogeneous set of five types of data sources.

A third class of fusion algorithms has been developed for target or event detection. Although related to the previous schemes, now the focus is to optimize the set of parameters that must be estimated to support detection. For example, Dewasurendra *et al.* [20] applied the Dempster–Shafer evidence theory to process information from multiple sensor modalities. This Bayesian framework allows one to update incoming data and consider the reliability of the different sensors. A related problem is discussed by Zhu *et al.* [21]. The data acquired by different sensors are first compressed at each sensor before being transmitted to a central processor. They derive an application-specific compression matrix, which minimizes the estimate error variance of a parameter, related to the measurements by a linear model.

Our approach, presented in this paper, relates to all classes of algorithms. First, we recover the underlying manifold of each input sensor separately. In particular, we apply the density invariant embedding (Section III-A) that overcomes the fundamental obstacle posed by the sampling density variation over the different sensors. Secondly, in Section VI, we derive and apply an application-specific feature selection approach to improve recognition accuracy. This combination of embedding coordinates and feature selection should be useful for many of the problems listed above.

III. Diffusion Embeddings and Extensions

The diffusion framework provides two essential computational tools for pattern recognition: i) density invariant embeddings (Section III-A) and ii) the extension of a given embedding to a new set of points (Section III-B). The first tool paves the way for a canonical embedding, invariant to the properties of a particular manifestation of a data source; the second allows the extension of the knowledge from a learning set to a test set. A broader applicative view and mathematical analysis can be found in [6]. Two quantitative approaches for the selection of the kernel bandwidth ε are presented in Section III-C.

A. Spectral and Density-Invariant Embeddings

Given a set $\Omega = \{\mathbf{x}_i\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^d$ of data points, we start by constructing a weighted symmetric graph in which each data point x_i corresponds to a node. In applications where a distance $d(\cdot,\cdot)$ already exists on the data, it is customary to weight the edge between $\mathbf{x}_i$ and $\mathbf{x}_j$ by $w(\mathbf{x}_i,\mathbf{x}_j) = \exp(-d(\mathbf{x}_i,\mathbf{x}_j)^2/\varepsilon)$, where $\varepsilon > 0$ is a scale parameter (kernel bandwidth).

We induce a random walk on the data set Ω by computing the row-normalized Markov matrix M

$$M(\mathbf{x}_i,\mathbf{x}_j) = \frac{w(\mathbf{x}_i,\mathbf{x}_j)}{\sum\limits_{\mathbf{x}_j\in\Omega} w(\mathbf{x}_i,\mathbf{x}_j)}.$$

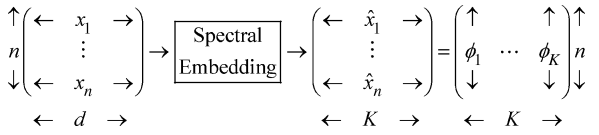


Fig. 2. Embedding the dataset. Given a set of n samples $\{\mathbf{x}_i\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^d$ we compute K eigenvectors, each of length n. Each eigenvector is a spectral embedding coordinate, and the embedding of each sample is given by $\{\widehat{\mathbf{x}}_i\}_{i=1}^n$, $\widehat{\mathbf{x}}_i \in \mathbb{R}^K$.

The embedding coordinates for the set Ω are given by the right K eigenvectors and eigenvalues of M

$$M\boldsymbol{\phi}_k = \lambda_k \boldsymbol{\phi}_k. \tag{III.1}$$

Denote the embedding $\widehat{\Omega} = \{\widehat{\mathbf{x}}_i\}_{i=1}^n, \widehat{\mathbf{x}}_i \in \mathbb{R}^K$. $\widehat{\Omega}$ consists of $\{\boldsymbol{\phi}_k\}_{k=1}^K$ spectral embedding vectors. Each embedding vector $\boldsymbol{\phi}_k$ is of length n and is the $k'th$ spectral coordinate. Thus, the i'th entry in $\boldsymbol{\phi}_k$ is the k'th spectral coordinate of the sample $\mathbf{x}_i$. This is depicted in Fig. 2.

Density-invariant embedding [4], [6] is essential to our application. Its focal point is to make the embedding reflect only the geometry of the data and not its density. Classical eigenmap methods [22], [23] provide an embedding that confounds information about the density of sample points and the manifold geometry, so the embedding coordinates heavily depend on the density of the data points. To remove the influence of the density of the data points, we renormalize the Gaussian edge weights $w_\varepsilon(\cdot,\cdot)$ with an estimate of the density and then compute the Graph–Laplacian embedding.

B. Out of Sample Extension

In most recognition tasks, it is essential to extend the low-dimensional representation computed on a training set to new samples. Let $\Omega = \{\mathbf{x}_i\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^d$ be a dataset; $\widehat{\Omega} = \{\widehat{\mathbf{x}}_i\}_{i=1}^n$, $\widehat{\mathbf{x}}_i \in \mathbb{R}^K$ its spectral embedding; and $\{\boldsymbol{\phi}_k\}_{k=1}^K$ the corresponding diffusion embedding vectors, either density-invariant or not. We present an approach for computing the embedding of a *new* point, $\mathbf{y} \in \mathbb{R}^d$, $\mathbf{y} \notin \Omega$, without having to embed the superset $\{\Omega, \mathbf{y}\}$. The embedding is extended to $\mathbf{y}$ by extrapolating the spectral embedding vectors $\{\boldsymbol{\phi}_k\}_{k=1}^K$ to $\mathbf{y}$ using the Nyström extension method [24]. Following (III.1), the spectral embeddings $\{\boldsymbol{\phi}_k\}_{k=1}^K$ are the eigenvectors of a positive semidefinite (psd) matrix, and the Nyström extension extrapolates the eigenvectors of psd kernels (matrices) to new points. Thus, by extending the K embedding vectors to the point $\mathbf{y}$, we drive $\widehat{\mathbf{y}} \in \mathbb{R}^K$ that is the spectral embedding of $\mathbf{y}$.

For ease of presentation, we apply the extension scheme to a Gaussian kernel, while other psd kernels can be used *mutatis mutandis*. Let $\sigma > 0$ be a scale of extension and consider the eigendecomposition of a Gaussian kernel of width σ on the training set Ω

$$\mu_k \boldsymbol{\phi}_k^j = \sum_i e^{-\|\mathbf{x}_j - \mathbf{x}_i\|^2/\sigma^2} \boldsymbol{\phi}_k^i, \quad \mathbf{x}_j \in \Omega$$

where $\boldsymbol{\phi}_k^i$ is the i'th entry of the vector $\boldsymbol{\phi}_k$.

As the kernel can be evaluated at any point in $\mathbb{R}^d$, it is possible to take any $\mathbf{y} \in \mathbb{R}^d$ in the right-hand side of this identity. This yields the following definition of the Nyström extension of ϕ_k from Ω to $\mathbb{R}^d$

$$\boldsymbol{\phi}_k^* \triangleq \frac{1}{\mu_k} \sum_i e^{-\|\mathbf{y} - \mathbf{x}_i\|^2/\sigma^2} \boldsymbol{\phi}_k^i, \quad \mathbf{y} \in \mathbb{R}^d \tag{III.2}$$

where $\boldsymbol{\phi}_k^*$ is the *extended* value of the eigenvector $\boldsymbol{\phi}_k$ at the point $\mathbf{y}$. By applying (III.2) to the K eigenvectors $\{\boldsymbol{\phi}_k\}_{k=1}^K$, we derive $\widehat{\mathbf{y}}$, the embedding of $\mathbf{y}$.

C. Kernel Bandwidth Selection

In this section, we discuss the selection of the kernel bandwidth parameter ε, used in Section III-A to derive the spectral embedding. For that, we provide two data-driven approaches. The first, suggested by Singer *et al.* [25] and denoted by the Singer measure (SM), is based on an asymptotic analysis of the discrete Markov walk induced by the spectral embedding on the dataset [26]. By analyzing the discretization error of the discrete Markov walk, it is shown [25] that the following functional:

$$SM(\varepsilon) = \sum_{i,j} w_{ij} = \sum_{i,j} \exp\left(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\varepsilon\right)$$

can be used to characterize the discretization error. The *sweet spot* of this curve, in terms of discretization error, is the interval of ε for which the curve is linear in a log-log plot of $SM(\varepsilon)$.

The second approach, denoted the *max-min measure,* is based on the notion that spectral embeddings agglomerate local parameterizations of the data manifold to derive a global parametrization. More accurately, ε should be chosen such that the kernel describes the infinitesimal connectivity of the data set. Thus, we choose ε to be as small as possible while maintaining its local connectivity. In practice, we use the following local connectivity measure:

$$H_i(\varepsilon) = \min_j \left(\|\mathbf{x}_i - \mathbf{x}_j\|^2\right), \quad i \neq j. \tag{III.3}$$

Thus, $H_i(\varepsilon)$ is the minimal distance connecting the data point x_i to its neighborhood. The kernel bandwidth is then set as

$$\varepsilon = C \max\left(H(\varepsilon)\right) \tag{III.4}$$

where C is typically in the range of two to three. We provide experimental validation of both approaches in Section VII-B.

IV. Multisensor Fusion

With the background in place, we now present the proposed data-fusion scheme. For the sake of clarity, we restrict our discussion to the case of two input channels; the extension to an arbitrary number of channels is straightforward and is discussed later in this section.

Suppose we are given two sets of measurements Ω_1, Ω_2 related to a single phenomenon $\Omega = \{x_1, \ldots, x_n\}$, where $\Omega_1 = \{\mathbf{x}_1^1, \ldots, \mathbf{x}_n^1\}$, $\mathbf{x}_i^1 \in \mathbb{R}^{d_1}$, and $\Omega_2 = \{\mathbf{x}_1^2, \ldots, \mathbf{x}_n^2\}$, $\mathbf{x}_i^2 \in \mathbb{R}^{d_2}$. We seek to fuse Ω_1 and Ω_2 by computing a unified low-dimensional representation $\widehat{\Omega} = \{\widehat{\mathbf{z}}_i\}_{i=1}^n$. It is assumed that Ω_1 and Ω_2

are aligned, implying that $\mathbf{x}_i^1$ and $\mathbf{x}_i^2$ relate to the same epoch of $\mathbf{x}_i$.

We start by computing the Laplace–Beltrami embedding (LBE) of Ω_1 and Ω_2, denoted $\widehat{\Omega}_1 = \{\widehat{\mathbf{x}}_i^1\}_{i=1}^n$, $\widehat{\mathbf{x}}_i^1 \in \mathbb{R}^{K_1}$ and $\widehat{\Omega}_2 = \{\widehat{\mathbf{x}}_i^2\}_{i=1}^n$, $\widehat{\mathbf{x}}_i^2 \in \mathbb{R}^{K_2}$, respectively. Each embedding reflects the *geometry* of the data as viewed by that sensor. To combine these representations into a unified representation $\{\widehat{\mathbf{z}}_i\}_{i=1}^n$, we compose them to form

$$\widehat{\mathbf{z}}_i = \begin{pmatrix} \widehat{\mathbf{x}}_i^1 & \widehat{\mathbf{x}}_i^2 \end{pmatrix}^T. \tag{IV.1}$$

$\widehat{\Omega}$ is made of n vectors of dimension $K_1 + K_2$.

Given L input channels $\{\Omega_l\}_{l=1}^L$, one would compute $\{\widehat{\Omega}_l\}_{l=1}^L$, the embeddings of the L input channels, and combine the embedding coordinates

$$\widehat{\mathbf{z}}_i = \begin{pmatrix} \widehat{\mathbf{x}}_i^1 & \dots & \widehat{\mathbf{x}}_i^L \end{pmatrix}^T. \tag{IV.2}$$

As before, each input sensor might have a different embedding dimensionality K_l.

A. Discussion

The geometrical embedding approaches (CCA and ours) assume that there exists a *common low-dimensional manifold*. CCA assumes it to be linear and proposes an optimal numerical scheme for its recovery, while our approach allows nonlinear manifolds. In contrast, MI implicitly assumes a *common statistical model*. These assumptions are application driven and axiomatic. For instance, there is a gamut of works on registering multisensor medical images by MI [10]. None of them proved that there exists a common probability distribution among the multimodality images. We assume its existence due to their perceptual similarity, overlooking the dissimilar appearance. We also assume, due to the setup of the acquisition process, that the input images capture the same body parts. If these assumptions do not hold, the registration will fail. The same applies to CCA, but there the common latent representation is a linear manifold. If it fails, the CCA will provide inferior results (see Section VII).

Another issue to consider is the numerical stability and accuracy. The CCA is based on eigendecomposition that is numerically stable and recovers the common linear manifold *optimally*. But, it is susceptible to the "curse of dimensionality" due to the estimation of the covariance matrix [13]. Our scheme assumes a common nonlinear manifold and is applicable for high-dimensional data. It also handles the sampling density issue. The estimation of MI of high-dimensional data sources is ill-posed due to the curse of dimensionality, and even for low-dimensional data, one has to resolve certain implementation details (density estimation scheme, interpolation, etc.).

V. Group Recognition in Diffusion Spaces

The mathematical foundations laid in the previous sections pave the way for a data-classification scheme able to handle *group objects*. Each group object consists of a set of high-dimensional samples. Most state-of-the-art classifiers, such as SVM and NN, classify a *single* high-dimensional sample at a time, and there is no notion of a group. In the context of this paper, we denote each such group as a *signal* $S = \{\mathbf{x}_i\}_{i=1}^{|S|}$. For instance, a *signal* S might be the utterance of a digit that consists of a set of high-dimensional *samples* (vectors), each sample $\mathbf{x}_i \in \mathbb{R}^d$ being a temporal window. Given that there are C different classes of signals, we derive a recognition scheme able to classify an input *test signal* $T = \{\mathbf{x}_i^T\}_{i=1}^{|T|}$ to a particular class.

The analysis of high-dimensional signals, such as audio and video signals, is of particular interest and utilizes low-dimensional embeddings as discussed in the previous sections. In particular, the spectral fusion scheme allows us to represent multiple heterogeneous sources by a unified spectral embedding in $\mathbb{R}^K$, extending the application of those techniques for representing the input as a single channel. This implies that we can utilize the same classifier for both single (audio, video) and multisensor (audio/visual) data. We now develop this into our group recognition scheme, which consists of three phases: learning the data manifold, learning the training signals, and recognition. Each is considered in turn.

A. Learning the Data Manifold of the Input Sensors

We start by learning the manifold of the high dimensional *samples* using spectral embedding. The manifold is learned by computing the embedding $\widehat{\Omega} = \{\widehat{\mathbf{x}}_i\}_{i=1}^n$, $\widehat{\mathbf{x}}_i \in \mathbb{R}^K$ of a set of samples $\Omega = \{\mathbf{x}_i\}_{i=1}^n$, $\mathbf{x}_i \in \mathbb{R}^d$; thus it should span the space of the input source. For instance, a set corresponding to the speech manifold of a speaker is acquired by having the subject read aloud an article. We thus assume that the article contains all possible phoneme and mouth positions.

The embedding $\widehat{\Omega}$ will be later used to efficiently compute the embeddings of other input samples via out-of-sample-extension (Section III-B). When the input dataset consists of a single sensor, one can use either the graph Laplacian embedding (GLE) or the density-invariant LBE. But, when dealing with a multisensor dataset, one has to use the LBE to fuse the input sensors.

B. Learning the Training Signals

In this phase, the signals are known to consist of C classes $c = 1, 2, \dots, C$; for our speech-recognition example, $C = 10$ corresponds to the digits '0', ..., '9'. We are also given the sample learning set Ω and its embedding $\widehat{\Omega}$.

We are given C sets of training signals $\{A_c\}_{c=1}^C$

$$A_c = \{S_i^c\}_{i=1}^{|A_c|} \tag{V.1}$$

where $S_i^c = \{\mathbf{x}_j^{c,i}\}_{j=1}^{|S_i^c|}$ is a training signal. This hierarchy is depicted in Fig. 3.

We start by computing the embedding of the training signals. The embedding of a signal S_c^i is given by $\widehat{S}_i^c = \{\widehat{\mathbf{x}}_j^{c,i}\}_{j=1}^{|S_i^c|}$, the embedding of its samples. $\widehat{S}_i^c$ has the same number of samples (vectors) as S_i^c, each being an embedding of dimension K. We denote by $\widehat{A}_c = \{\widehat{S}_c^i\}_{i=1}^{|A_c|}$ the embedding of a training set, given by the embeddings of its training signals. The embedding hierarchy is also depicted in Fig. 3. $\{\widehat{A}_c\}_{c=1}^C$ is computed by extending the embedding $\widehat{\Omega}$ of the learning set to each sample $\mathbf{x} \in \{A_c\}_{c=1}^C$.

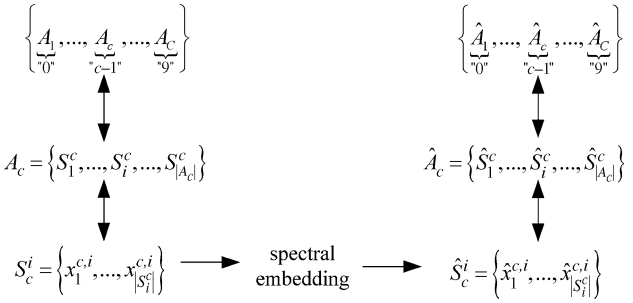


Fig. 3. The data hierarchy of the group-recognition problem. Each item to recognize is a signal S_c^i consisting of high-dimensional samples. For recognition purposes, we are given $\{A_c\}_{c=1}^C$ sets of training signals.

C. Test Signal Recognition

We conclude with the *signal recognition step*, where, given a test signal $T = \{\mathbf{x}_i^T\}_{i=1}^{|T|}$, $\mathbf{x}_i^T \in \mathbb{R}^d$, we aim to identify its corresponding class c^*. For that, we compute $\widehat{T}$ the embedding of T by way of spectral extension as before. Simply put, the spectral extension allows us to position both the training and test sets within the same K-dimensional diffusion based coordinates system.

Given the embeddings of the *Learning* and *Test* sets, we require a group object similarity measure able to quantify the notion of distance between sets of points in $\mathbb{R}^d$. For that, we utilize the symmetric Hausdorff distance (SHD) . Given two points sets $S_1 \in \mathbb{R}^d$ and $S_2 \in \mathbb{R}^d$, the SHD $d_H(S_1, S_2)$ is given by

$$d_H(S_1,S_2) = \max\left\{ \max_{\mathbf{x}_2 \in S_2} \min_{\mathbf{x}_1 \in S_1} \left\{ \|\mathbf{x}_1 - \mathbf{x}_2\|^2 \right\}, \max_{\mathbf{x}_1 \in S_1} \min_{\mathbf{x}_2 \in S_2} \left\{ \|\mathbf{x}_1 - \mathbf{x}_2\|^2 \right\} \right\}. \quad \text{(V.2)}$$

The SHD is a min-max distance that does not require the matching or alignment of the sets S_1 and S_2. It is fast to compute and assumes that both sets are given in the same coordinates system. It naturally handles nonlinear elastic deformations between the sets. As such, it is suitable for the group object recognition problem in general and the audiovisual task in particular.

It follows by (V.2) that the SHD is determined by a single pair of samples $\mathbf{a} \in S_1$ and $\mathbf{b} \in S_2$ that are the furthest away, among all pairs of nearest neighbors in S_1 and S_2. This observation paves the way for the iterative SHD feature selection scheme presented in Section VI. Given the SHD that quantifies the distance between $\widehat{T}$ and the embedding of the training sets $\widehat{A}_c$, one can train and apply a variety of classification schemes, such as SVM, LDA, and K-NN classifiers.

In order to exemplify the effectivity of sensor fusion process, we applied the 1NN classifier

$$c^* = \arg \min_{\substack{\overline{c}=1\ldots C \\ i=1\ldots|A_{\overline{c}}|}} d_H\left(\widehat{T}, S_i^{\overline{c}}\right). \quad \text{(V.3)}$$

Simply put, we look for the training signal $S_i^{c^*}$ whose embedding is the closest to $\widehat{T}$ and classify T as $T \in c^*$.

VI. Spectral Features Selection

Our recognition scheme is based on computing the SHD between sets of points. In this section, we derive a feature selection scheme that improves the recognition rate by maximizing the SHD classification margin. We emphasize that other common feature selection approaches [3], [27], such as discriminant analysis (LDA, QDA) and SVM, are inapplicable to group recognition, since they maximize the recognition margin of a *single* sample, and not that of a *group* (signal).

Let S^c be a signal belonging to the class c; its SHD classification margin $\Delta(S^c)$ is given by

$$\Delta(S^c) \triangleq \min_{\substack{\overline{c}=1\ldots C \\ \overline{c}\neq c}} \min_{i=1\ldots|A_{\overline{c}}|} d_H\left(S^c, S_i^{\overline{c}}\right) - \min_{i=1\ldots|A_c|} d_H\left(S^c, S_i^c\right). \quad \text{(VI.1)}$$

$\Delta(S^c)$ is the difference in SHD scores between the best false classification of S^c [first term on the right-hand side of (VI.1)] and the best true classification [second term on the right-hand side of (VI.1)]. We aim to have $\Delta(S^c) \gg 0$, as the signal S^c is expected to be closer to the signals in c rather to those in $\overline{c} \neq c$.

Our supervised approach to feature selection utilizes two given training sets of signals $A_F^c \notin c$ and $A_T^c \in c$. We aim to compute a vector of *class-specific* weights $\mathbf{w}_c \in \mathbb{R}^K$, $\sum_{m=1}^K \mathbf{w}_m^c = 1$, such that it maximizes the *weighted SHD margin*

$$\Delta_w(S^c) \triangleq \min_{S \in A_F} d_H(W_C S^c, W_C S) - \min_{S \in A_T} d_H(W_C S^c, W_C S) \quad \text{(VI.2)}$$

where W_C is a $K \times K$ diagonal matrix with $\mathbf{w}_c$ as its main diagonal. By $W_C S$, we imply that each sample $\mathbf{x} \in S, \mathbf{x} \in \mathbb{R}^K$, is pointwise weighted by $\mathbf{w}_c$.

The optimal weight vector $\mathbf{w}_c$ is given by

$$\mathbf{w}_c(S^c) = \arg\max_{\mathbf{w}} \Delta_w(S^c). \quad \text{(VI.3)}$$

For now, we show how to maximize the margin for a particular training sample S^c; we will later generalize the scheme for a set of training samples $\{S_i^c\}$.

Solving (VI.3) directly can prove difficult, since, for each weight vector $\mathbf{w}_c$, there could be different nearest neighbors within the computation of the SHD. Hence, we propose the following iterative scheme. We start by assuming a uniform weight vector $\mathbf{w}_c$, $\mathbf{w}_m^c = 1/K, \forall m$. Given $\mathbf{w}_c$, (VI.3) can be simplified by searching over A_T^c and A_F^c for the signals closest to S^c

$$S_F = \arg\min_{S \in A_F^c} d_H(W_C S^c, W_C S)$$

and

$$S_T = \arg\min_{S \in A_T^c} d_H(W_C S^c, W_C S).$$

We can then find the pairs of samples $(\mathbf{a}^{Fa} \in S^c, \mathbf{b}^{Fa} \in S_F)$ and $(\mathbf{a}^{Tr} \in S^c, \mathbf{b}^{Tr} \in S_T)$ that determine the SHD margin. Inserting those into (VI.2), we get

$$\begin{aligned}\Delta_w(S^c) &= \sum_{m=1}^{K} (\mathbf{w}_m^c)^2 \left(\mathbf{a}_m^{Fa} - \mathbf{b}_m^{Fa}\right)^2 \\ &\quad - \sum_{m=1}^{K} (\mathbf{w}_m^c)^2 \left(\mathbf{a}_m^{Tr} - \mathbf{b}_m^{Tr}\right)^2 \\ &= \sum_{m=1}^{K} (\mathbf{w}_m^c)^2 \boldsymbol{\Delta}_m(S^c) \qquad \text{(VI.4)}\end{aligned}$$

where $\boldsymbol{\Delta}(S^c)$ is the *margin per feature*.

Given that $\sum_{m=1}^{K} \mathbf{w}_m^c = 1$, one is tempted to adopt a greedy approach to maximize the margin $\Delta(S^c)$ in (VI.4) such that a single element is set to $\mathbf{w}_{m_0}^c = 1$ and the rest are set to zero, where

$$m_0 = \arg\max_m \boldsymbol{\Delta}_m(S^c).$$

In practice, this approach proved to be too greedy, so we only used $\boldsymbol{\Delta}(S^c)$ as a ranking measure. Then, at each iteration, we choose the P features with the largest margins $\boldsymbol{\Delta}(S^c)$. Finally, to gain robustness, we compute the *average margin per feature* $\boldsymbol{\Delta}_c$ by averaging the margins $\boldsymbol{\Delta}_m(S^c)$ of all of the signals in A_T^c

$$\boldsymbol{\Delta}_c = \frac{1}{|A_T^c|} \sum_{S^c \in A_T^c} \boldsymbol{\Delta}(S^c).$$

As before, we choose the P features with the largest margins. Given the updated weight vector $\mathbf{w}_c$, we reiterate the process. This procedure is repeated for all C classes, each having its own weight vector $\mathbf{w}_c$. We found this scheme to converge within two to three iterations and provide improved recognition rates for small numbers of features. More experimental details are provided in Section VII.

VII. Experimental Results

Our sensor fusion scheme was put to the test by applying it to audio-visual speech recognition. We compare the proposed multisensor analysis scheme to visual-only and audio-only recognition and show an accuracy improvement. This application was chosen as it is a classical example of multisensor-based recognition. In Section VII-E, we compare our fusion scheme to principal component analysis (PCA) and CCA because these are considered state-of-the-art techniques [13], [14]. Our performance comparison demonstrates the power of the embedding coordinates; that they can be used to extend the above techniques demonstrates their potential power.

A. Data Acquisition and Preprocessing

We recorded several grayscale movies depicting the lips of a subject reading a text in English together with the audio track. The recording was made with a regular DV camera, and the audio was sampled at 32 KHz.

Each video frame was automatically cropped into a rectangle of size 140×110 around the lips and was considered as a point in $\mathbb{R}^{140\times110}$. The frames were normalized to have zero mean and unit variance.

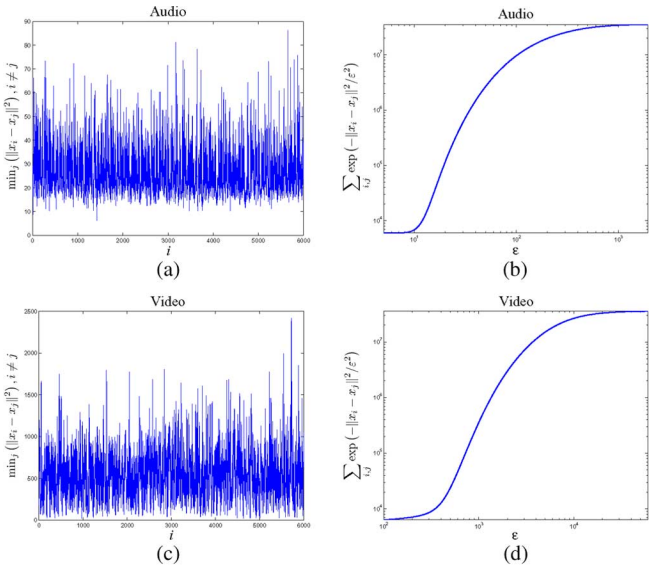


Fig. 4. Setting the kernel bandwidth ε. The maximal minimal distance per sample is shown in (a) and (c) for the audio and visual data, respectively. It is common to set ε to be twice as large as the maximal minimal distance. In (b) and (d), we compute the measure suggested by Singer [25]. The interval of ε for which the measure is linear is the preferred one.

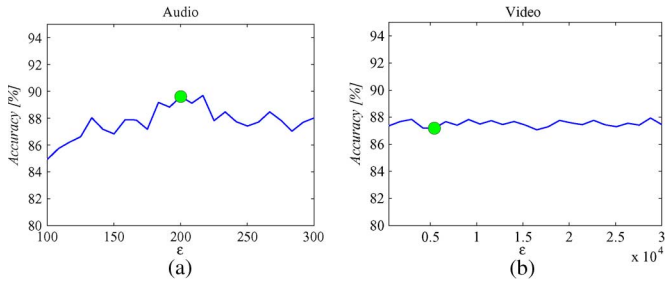


Fig. 5. The recognition accuracy using a single sensor versus the kernel bandwidth ε. The green dot marks the ε value based on the "max-min" criterion. Varying ε over a wide range does not change the recognition accuracy significantly.

The audio signal was broken up into overlapping time-windows centered at the beginning of each video frame. Since the video is sampled at 25 frames per second, we formed audio windows 80 ms long (equal in duration to two video frames) and reduced the frame-splitting artifacts by smoothing with a bell-shaped function. We computed the log of the magnitude of the fast Fourier transform (FFT) of each window, retaining only half of it, due to the conjugate symmetry of the FFT. We then computed the discrete cosine transform of the result. This produces a feature vector in $\mathbb{R}^{1280}$. To reduce the volume of the dataset, we averaged the vector into 256 uniform bins.

Using the conventions of Section IV, the set $\Omega_{\mathrm{A}} \in \mathbb{R}^{256}$ corresponds to the audio samples, while $\Omega_{\mathrm{V}} \in \mathbb{R}^{15400}$ corresponds to the set of video frames. Note that in this setup, Ω_{A} and Ω_{V} are naturally aligned and contain the same number of points.

Following Section V, the *manifold learning* dataset consisted of 6000 video frames (and as many audio windows), corre-

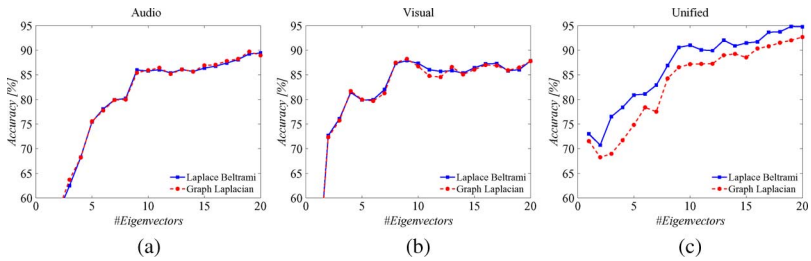


Fig. 6. Digits recognition results for Laplace–Beltrami (density invariant) and graph Laplacian embeddings. (a) and (b) Recognition accuracy of the single sensor classifiers. (c) The recognition results of the unified sensor scheme.

sponding to the speaker reading a press article. We will refer to these data as *text data*. In order to acquire the *training* and *test* sets of digits, we asked the subject to repeat each digit "zero," "one," …, "nine" 50 times. We will refer to these data as *digits data*. A typical digit consists of 20–30 samples.

B. Data Fusion and Kernel Bandwidth Selection

The audio and visual *text data* sets were embedded using the LBE discussed in Section III-A. In order to set the kernel bandwidth ε, we applied both the approaches presented in Section III-C. Fig. 4 depicts both bandwidth selection measures for the audio and visual datasets. Based on Fig. 4(a) and (c), and using a multiplier of $C = 2$ in (III.4), we set $\varepsilon_{\mathrm{A}} = 200$ for the audio data and $\varepsilon_{\mathrm{V}} = 5000$ for the video data. Examining the other measure in Fig. 4(b) and (d), it is evident that these values coincide with the linear interval of the curve.

All of the spectral embeddings were computed using the Gaussian kernel $\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/\varepsilon)$, and the test and training sets were created by randomly splitting the *digits data* into two parts. We denote the ratio of the learning set to the digits data as the *learning ratio* (LR). Following the second step in Section V, we extended the spectral embeddings from the text data to the digits data. These embeddings were computed separately for Ω_{A} and Ω_{V} and denoted $\widehat{\Omega}_{\mathrm{A}}$ and $\widehat{\Omega}_{\mathrm{V}}$, respectively. The unified embedding is then given by $\widehat{\Omega}_{\mathrm{AV}}$. Denote the corresponding Hausdorff distance based classifiers H_{A}, H_{V}, and H_{AV} for the audio, visual, and unified classifiers, respectively. We emphasize that the same embeddings $\widehat{\Omega}_{\mathrm{A}}$ and $\widehat{\Omega}_{\mathrm{V}}$ were used to derive all classifiers.

We then applied H_{A} and H_{V} to study the sensitivity of the recognition rate to the choice of the bandwidth ε. The results are reported in Fig. 5, where we measured the recognition rate of each channel separately. The recognition rate is averaged over all ten digits. The recognition is robust to changes of the bandwidth. For these tests, we used LR$= 0.5$, LBE, and $N_{\mathrm{ev}} = 20$ diffusion coordinates per channel, such that $\widehat{\Omega}_{\mathrm{A}} \in \mathbb{R}^{20}$ and $\widehat{\Omega}_{\mathrm{V}} \in \mathbb{R}^{20}$. The N_{ev} spectral features were chosen according to their eigenvalues' magnitude. (The first eigenvector taken is the one corresponding to the largest eigenvalue.)

C. Density Invariant Embedding

In this section, we study the use of density invariant spectral embeddings for the single and multisensor cases. We start by computing the approximate densities of Ω_{A} and Ω_{V}. Recalling that the density is positive and depends on the dynamic range

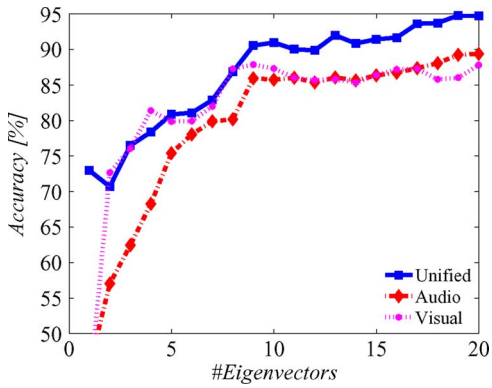


Fig. 7. Digits recognition versus the number of embedding coordinates for audio, visual, and audio-visual data. All channels were embedded using a density invariant embedding. The unified audio-visual scheme significantly outperforms the single sensor recognition.

of the data set, we use the following measure for evaluating the density uniformity:

$$U(\Omega) = std\left(D(\Omega)\right)/mean\left(D(\Omega)\right) \qquad \text{(VII.1)}$$

where D is the density of each dataset. The density was computed by kernel-based density estimation, using Gaussian kernels with the bandwidth computed in the previous section. The density comes out to be $U_{\mathrm{V}}(\Omega) = 0.0456$ and $U_{\mathrm{A}}(\Omega) = 0.2531$. It follows that the density of the visual channel is more uniform than that of the audio channel. In terms of manifold learning, this implies that the visual signal was able to uniformly span the space of speech states.

Fig. 6 studies the influence of the density invariant embedding on the classification accuracy. We report the classification results for both the single and multisensor cases, where the data are embedded using both the LBE and GLE. By that we aim to experimentally verify the conjecture used throughout this paper: the notion that the audio and visual data are different manifestations of a *common underlying low-dimensional process*. Using the LBE, both Ω_{A} and Ω_{V} are represented in their common diffusion coordinates system.

Fig. 6(a) and (b) compares the classification results in the single channel case (Ω_{A} and Ω_{V}) using the LBE and GLE, with LR$= 0.5$ and $N_{\mathrm{ev}} = 20$. There is no significant difference in the recognition accuracy of the LBE and GLE. In contrast, Fig. 6(c) depicts the classification results for our proposed sensor fusion scheme, where the LBE outperforms the GLE.

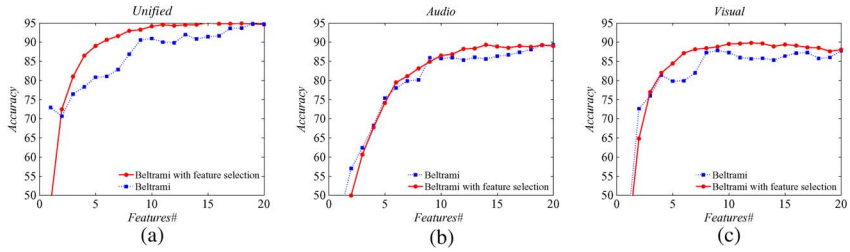


Fig. 8. Feature selection results. The feature selection scheme outperforms the naive selection approach for all classifiers.

D. Unified Recognition Results and Feature Selection

Given the classifiers H_{A}, H_{V}, and H_{AV}, we compare their recognition accuracies in Fig. 7. In the multisensor case, we use N_{ev} embedding vectors from each input channels ($2N_{\mathrm{ev}}$ embedding vectors overall). For a small number of coordinates $N_{\mathrm{ev}} \leq 8$, the unified classifier performs as well as H_{V}, which is the best of H_{A} and H_{V}. But, for $N_{\mathrm{ev}} > 8$, we begin to witness the fusion effect, as the proposed fused classifier H_{AV} is able to outperform each of the single sensor classifiers H_{A} and H_{V}. In particular, the recognition rate of H_{V}, the best of the single sensor classifiers, saturates around 88%, while the unified one H_{AV} achieves an accuracy of up to 95%. Adding more diffusion coordinates ($N_{\mathrm{ev}} > 20$) to H_{A} and H_{V} did not improve the classification accuracy.

We then studied the feature selection scheme introduced in Section VI, where we use the same number of coordinates N_{ev} for all classifiers. The results are shown in Fig. 8, where the feature selection performed well for all classifiers, being superior to the common approach of selecting the spectral coordinates according to their eigenvalues.

As for computational complexity, the proposed scheme is asymmetric by nature. Namely, most computations are done offline, computing the manifold embedding vectors, so that the recognition step is fast. The typical running time for the preprocessing of 6000 audio and visual frames and the computation of the spectral embeddings was approximately 3 min. We run our Matlab implementation on an Intel Core2 Duo running at 2.20 GHz with 4 GB of memory. The recognition of a single digit, consisting of 40 samples on average, took an average of 0.5 s.

E. Comparison to Canonical Correlation Analysis and Principal Component Analysis

We compared the proposed unified recognition approach to linear dimensionality reduction schemes based on PCA. Those were found to be useful in previous works [13], [14]. The results are depicted in Fig. 9, and the maximal recognition rates for each scheme are reported in Table I. First, we applied CCA directly to the audio/video features. The recognition rate was quite low, topping at 66.7% compared to the 94.8% achieved by the proposed scheme. The CCA aims to recover the two most correlated linear components, one from each sensor. Yet, in high-dimensional signals, such as our audio/video data, this assumption might prove too restrictive, as the signals are related by nonlinear transforms. We also applied PCA to both sensors separately and appended the embeddings, the same way we fused

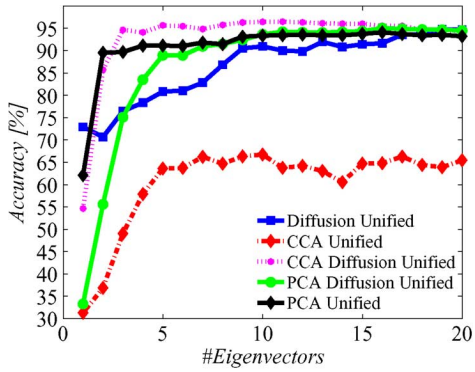


Fig. 9. Recognition rates for the different sensor fusion schemes. The maximal retrieval rates are reported in Table I.

TABLE I
RETRIEVAL RATES FOR THE DIFFERENT SCHEMES

	Spectral	Spectral+PCA	Spectral+CCA	CCA	PCA
Rate	94.82%	95.1%	96.52%	66.79%	91.75%

the spectral embeddings in Section IV. The resulting recognition rate was up to 91.75%. Motivated by that, we applied the CCA and PCA to the spectral embeddings. The CCA improved the recognition rate significantly and achieved a recognition rate of 96.5%, the highest over all schemes. Applying the PCA to the spectral embeddings gained an accuracy of 95.1%.

This is attributed to two fundamental issues: first, the CCA recovers the most correlated *linear* structures within the data, while the diffusion framework recovers *nonlinear* structures. Secondly, Kidron *et al.* [13] show that the CCA is ineffective in analyzing high-dimensional data, as the estimation of the corresponding covariance matrix becomes statistically ill-posed. In [13], the issue was resolved by applying a sparsity prior, as the common manifold related to audio/visual events was sparse along the temporal axis. This is not the case in our data.

VIII. SUMMARY AND CONCLUSIONS

In this paper, we introduced and tested a multisensor data analysis scheme based on spectral embedding. The scheme was applied to audio-visual speech recognition. We embedded each data source separately and then appended the embeddings to derive the fused representation. The success of the scheme provides some evidence that data manifolds can be "composed" in support of recognition. Such compositions could emerge as a fundamental tool in the future.

Our recognition approach was explicitly based on the notion of *group recognition,* and an original contribution of this paper was the development of a corresponding novel feature selection scheme. This provides a data point to confirm the power of diffusion coordinates as a foundation for recognition algorithms and extends the basis for classifiers to work with them. Diffusion embeddings require the specification of a resolution or bandwidth parameter in the kernel, and we also detailed two data-driven approaches for the selection of it.

In summary, the experimental results show that the fusion effect is indeed achieved, and that multisensor-based recognition can be made superior to single-sensor classifiers. Our final result, that PCA analysis applied to the spectral embedding further improves the recognition rate, suggests that even more is possible. Perhaps this is because even linear techniques can yield improvements when the requisite nonlinear structure is captured by proper embeddings.

ACKNOWLEDGMENT

The authors would like to thank Associate Editor Prof. P. K. Varshney and the anonymous referees for their thorough and constructive questions and comments.

REFERENCES

[1] J. Driver, "Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading," *Nature*, no. 381, pp. 66–68, 1996.
[2] Y. Gutfreund, W. Zheng, and E. I. Knudsen, "Gated visual input to the central auditory system," *Science*, no. 297, pp. 1556–1559, 2002.
[3] T. Hastie, R. Ribshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer, 2002.
[4] R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker, "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps," *Proc. Nat. Acad. Sci.*, vol. 102, no. 21, pp. 7426–7431, May 2005.
[5] R. Coifman, S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S. Zucker, "Geometric diffusions as a tool for harmonics analysis and structure definition of data: Multiscale methods," *Proc. Nat. Acad. Sci.*, vol. 102, no. 21, pp. 7432–7437, May 2005.
[6] R. Coifman and S. Lafon, "Diffusion maps," *Appl. Comput. Harmon. Anal. (Special Issue on Diffusion Maps and Wavelets)*, vol. 22, pp. 5–30, Jul. 2006.
[7] A. Gueziec, X. Pennec, and N. Ayache, "Medical image registration using geometric hashing," *IEEE Comput. Sci. Eng. Mag. (Special Issue on Geometric Hashing)*, vol. 4, pp. 29–41, Oct.–Dec. 1997.
[8] H. Li, B. S. Manjunath, and S. K. Mitra, "A contour-based approach to multisensor image registration," *IEEE Trans. Image Process.*, vol. 4, pp. 320–334, Mar. 1995.
[9] R. Sharma and M. Pavel, "Registration of video sequences from multiple sensors," in *Proc. Image Registr. Workshop (NASA GSFC)*, 1997, pp. 361–366.
[10] P. Viola, I. Wells, and W. M. , "Alignment by maximization of mutual information," in *Proc. 5th Int. Conf. Comput. Vision*, Jun. 1995, pp. 16–23.
[11] U. Ozertem and D. Erdogmus, "Information regularized sensor fusion: Application to localization with distributed motion sensors," *J. VLSI Signal Process. Syst.*, vol. 49, no. 2, pp. 291–299, 2007.
[12] J. Ham, D. Lee, and L. Saul, "Semisupervised alignment of manifolds," in *Proc. 10th Int. Workshop Artif. Intell. Statist.*, Jan. 6–8, 2005, pp. 120–127.
[13] E. Kidron, Y. Schechner, and M. Elad, "Cross-modal localization via sparsity," *IEEE Trans. Signal Process.*, vol. 55, pp. 1390–1404, Apr. 2007.
[14] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Audiovisual synchronization and fusion using canonical correlation analysis," *IEEE Trans. Multimedia*, vol. 9, no. 7, pp. 1396–1403, 2007.
[15] S. Lafon, Y. Keller, and R. R. Coifman, "Data fusion and multi-cue data matching by diffusion maps," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 28, no. 11, pp. 1784–1797, 2006.
[16] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the smo algorithm," in *Proc. ACM 21st Int. Conf. Machine Learn. (ICML '04)*, New York, 2004, p. 6.
[17] F. Bach, "Exploring large feature spaces with hierarchical multiple kernel learning," in *Proc. NIPS*, 2008, pp. 105–112.
[18] G. R. Lanckriet, M. Deng, N. Cristianini, M. I. Jordan, and W. S. Noble, "Kernel-based data fusion and its application to protein function prediction in yeast," in *Proc. Pacific Symp. Biocomput.*, Jan. 2004, pp. 300–311.
[19] P. Pavlidis, J. Weston, J. Cai, and W. N. Grundy, "Gene functional classification from heterogeneous data," in *Proc. 5th Annu. Int. Conf. Comput. Mol. Biol.*, New York, 2001, pp. 242–248.
[20] D. Dewasurendra, P. Bauer, and K. Premaratne, "Evidence filtering," *IEEE Trans. Signal Process.*, vol. 55, pp. 5796–5805, Dec. 2007.
[21] Y. Zhu, E. Song, J. Zhou, and Z. You, "Optimal dimensionality reduction of sensor data in multisensor estimation fusion," *IEEE Trans. Signal Process.*, vol. 53, pp. 1631–1639, May 2005.
[22] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Comput.*, vol. 6, no. 15, pp. 1373–1396, Jun. 2003.
[23] M. Balasubramanian, E. L. Schwartz, J. B. Tenenbaum, V. de Silva, and J. C. Langford, "The Isomap algorithm and topological stability," *Science*, vol. 295, no. 5552, p. 7, 2002.
[24] C. Fowlkes, S. Belongie, F. Chung, and J. Malik, "Spectral grouping using the nyström method," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 26, no. 2, pp. 214–225, 2004.
[25] R. R. Coifman, Y. Shkolnisky, F. J. Sigworth, and A. Singer, "Graph laplacian tomography from unknown random projections," *IEEE Trans. Image Process.*, vol. 10, pp. 1891–1899, Oct. 2008.
[26] A. Singer, "From graph to manifold laplacian: The convergence rate," *Appl. Comput. Harmon. Anal.*, vol. 1, no. 21, pp. 135–144, 2006.
[27] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Machine Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

Yosi Keller received the B.Sc. degree from the Technion—Israel Institute of Technology, Haifa, in 1994 and the M.Sc. and Ph.D. degrees from Tel Aviv University, Tel Aviv, Israel, in 1998 and 2003, respectively, all in electrical engineering.

From 1994 to 1998, he was an R&D Officer in the Israeli Intelligence Forces. From 2003 to 2006, he was a Gibbs Assistant Professor with the Department of Mathematics, Yale University, New Haven, CT. He is a Senior Lecturer at the Electrical Engineering, Bar Ilan University, Israel. His research interests include graph-based data analysis, optimization, and spectral graph theory-based dimensionality reduction.

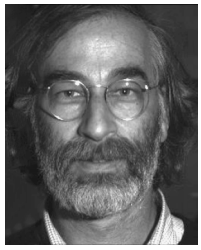
Ronald R. Coifman received the Ph.D. degree from the University of Geneva, Switzerland, in 1965.

He is the Phillips Professor of mathematics at Yale University, New Haven, CT. His research interests include nonlinear Fourier analysis, wavelet theory, singular integrals, numerical analysis and scattering theory, and new mathematical tools for efficient computation and transcriptions of physical data, with applications to numerical analysis, feature extraction recognition, and denoising.

Prof. Coifman is a member of the National Academy of Sciences and the American Academy of Arts and Sciences. He received the DARPA Sustained Excellence Award in 1996 and the 1999 Pioneer Award from the International Society for Industrial and Applied Mathematics. He received a National Medal of Science.

Stéphane Lafon received the B.Sc. degree in computer science from Ecole Polytechnique, France, the M.Sc. degree in mathematics and artificial intelligence from Ecole Normale Supérieure de Cachan, France, and the Ph.D. degree in applied mathematics from Yale University, New Haven, CT, in 2004.

He is a Software Engineer with Google, Inc. He was a Research Associate with the Mathematics Department, Yale University, in 2004–2005. His work with Google focuses on the design, analysis, and implementation of machine-learning algorithms. His research interests are in data mining, machine learning, and information retrieval.



Steven W. Zucker (F'88) is the David and Lucile Packard Professor at Yale University, New Haven, CT, where he is the Director of the Program in Applied Mathematics, Professor of computer science, Professor of biomedical engineering, and a Member of the Interdisciplinary Neuroscience Program. Before joining Yale in 1996, he was a Professor of electrical engineering at McGill University; Director of the Program in Artificial Intelligence and Robotics of the Canadian Institute for Advanced Research; and a Cofounder of the McGill Research Center for Intelligent Machines.

Prof. Zucker is a Fellow of the Canadian Institute for Advanced Research and (by)Fellow of Churchill College, Cambridge. He shared the Siemens Award (with M. Langer) at CVPR '97.