Cross-Modal Localization via Sparsity

Einat Kidron, Yoav Y. Schechner, and Michael Elad

***Abstract*—Cross-modal analysis is a natural progression beyond processing of single-source signals. Simultaneous processing of two sources can reveal information that is unavailable when handling the sources separately. Indeed, human and animal perception, computer vision, weather forecasting, and various other scientific and technological fields can benefit from such a paradigm. A particular cross-modal problem is *localization*: out of the entire data array originating from one source, localize the components that best correlate with the other. For example, auditory and visual data sampled from a scene can be used to localize visual events associated with the sound track. In this paper we present a rigorous analysis of fundamental problems associated with the localization task. We then develop an approach that leads efficiently to a unique, high definition localization outcome. Our method is based on canonical correlation analysis (CCA), where inherent ill-posedness is removed by exploiting sparsity of cross-modal events. We apply our approach to localization of audio-visual events. The proposed algorithm grasps such dynamic audio-visual events with high spatial resolution. The algorithm effectively detects the pixels that are associated with sound, while filtering out other dynamic pixels, overcoming substantial visual distractions and audio noise. The algorithm is simple and efficient thanks to its reliance on linear programming, while being free of user-defined parameters.**

***Index Terms*—Computer vision, cross-sensor fusion, multimedia, multimodal analysis, multisensor fusion, overfitting, regularization, stochastic analysis.**

I. Introduction

There is a growing interest in cross-modal analysis, where two different modalities are processed simultaneously. Such processing often involves comparisons of vector arrays, such as images. It may also use observation of the vectors over time, where mutual correlation is sought. Examples for temporal correlations of data arrays appear in various fields: in climatology [1], [2], dynamic weather phenomena in a certain place are correlated to synoptic meteorological data, acquired over time and in several locations; in economy [3], correlations are pursued between revenue performance of a market versus a large set of economic and social criteria; in medical research [4], [5], correlations are sought between body reaction to external stimuli, or between rates of contacting, or recovering from a certain disease versus lifestyle data (consumption of sugars, proteins, vitamins) and treatment parameters.

This paper deals with such cross-modal correlations. A particular task in this regard is *localization*: out of the entire data array originating from one source, localize the components that best correlate with the other. We perform a rigorous analysis of fundamental problems associated with this task. As it turns out, it is difficult to obtain high quality localization if only a few samples exist for each variable vector. In the context of the above-mentioned examples, such is the case when handling fast changing events; having too few temporal samples of meteorological events; or when only a few subjects participate in a medical test. In such scenarios, we show that the localization problem is ill-posed. As a case study for cross-modal correlation, in this work we focus on *visual motion* that is associated with *audio*. Nevertheless, the mathematical approach we develop here is of a general nature, and can be applied to other fields, such as those mentioned above.

Activity in audio-visual cross-modal analysis has various research aspects, including lip reading [6]–[8], analysis and synthesis of music from motion [9], audio filtering based on motion [10], source separation based on vision [11]–[15], and emotion recognition [16]. We note that physiological evidence and analysis of biological systems show that fusion of audio-visual information is used to enhance perception [17]–[19]. In this field, the localization task seeks to accurately *pinpoint visual features* (image pixels) that are associated with audio sources. These pixels should be distinguished from other moving objects. We do not limit the problem to talking faces or other specific classes of sources, but seek a general algorithm to achieve this goal. Some existing methods use several microphones, where stereo triangulation indicates the spatial location of the sources [20]–[23]. In contrast, we seek a sharp spatial localization of the sound source, using a single microphone (emulating monaural hearing) and a video stream. Moreover, we wish good localization performance, even if interfering sounds exist, unrelated to the desired object.

As indicated in Fig. 1, audio and visual data are inherently difficult to compare because of the huge dimensionality gap between these modalities. To overcome this, a common practice is to project each modality into a one-dimensional (1-D) subspace [8], [13], [15]. Thus, two 1-D variables represent the audio and the visual signals. Localization algorithms typically seek 1-D representations that best correlate [8], [12], [13]. However, as we show, this approach has a fundamental flaw. The projection of the visual data is controlled by many degrees of freedom. Hence, a substantial amount of data is necessary to reliably learn the cross-relationships. For this reason, some methods use a very aggressive pre-pruning of visual areas or features to reduce the

Manuscript received June 10, 2005; revised April 19, 2006. Y. Schechner is a Landau Fellow, whose work was supported by the Taub Foundation, and as Alon Fellow. This work was supported in part by the Israeli Science Foundation under Grant 315/04, and by the Ollendorff Minerva Center, Electrical Engineering Department at the Technion. The Minerva Center is funded through the BMBF. This work was also supported in part by the S. Langberg Research Fund. The associate editor coordinating the review of this paper and approving it for publication was Dr. Ercan E. Kuruoglu.

E. Kidron and Y. Y. Schechner are with the Department of Electrical Engineering, Technion-Israel Inst. Technology, Haifa 32000, Israel (e-mail: einat.kidron@gmail.com; yoav@ee.technion.ac.il).

M. Elad is with the Department of Computer Science, Technion-Israel Inst. Technology, Haifa 32000, Israel (e-mail: elad@cs.technion.ac.il).

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.2006.888095

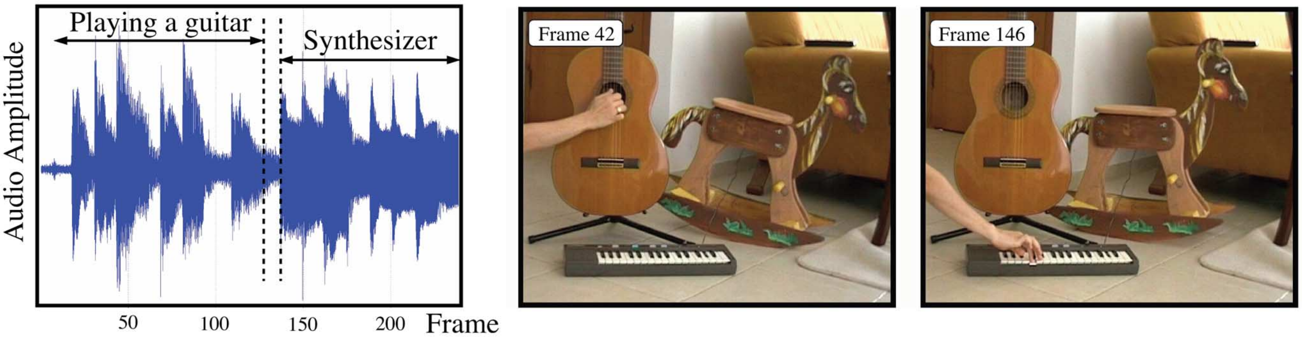


Fig. 1. The audio data is sequential, requiring $\mathcal{O}(10^4)$ samples/sec. Corresponding video frames are highly parallel (multipixel), requiring $\mathcal{O}(10^7)$ samples/sec. Pinpointing the sound source in the images by correlation requires dimensionality reduction of the visual signal. This reduction involves of too many degrees of freedom.

number of unknowns. Others consider acquisition of very long sequences to ensure sufficient data quantities. Those approaches result in a severe loss of either spatial or temporal resolutions, or both.

Audio-visual association can also be performed by optimizing the mutual information (MI) of the modal representations, as has been shown in the pioneering work by Fisher *et al.* [11], [24]. MI indicates cross-modal statistical dependency better than cross correlation does. However, estimating MI using Parzen windows is complex, and there is no guarantee for a unique solution, due to the nonconvexity of MI [25]. Moreover, such an approach suffers from the problem of insufficient data, as indicated above, just as methods that are based on cross-correlation.

The algorithm described here addresses these difficulties, while being based on canonical correlation analysis (CCA). It results in high spatio-temporal localization, and a unique solution. We exploit the fact that typically visual cues that correspond to audio sources are *spatially localized*, and thus *sparsity* of the solution is an appropriate prior. This makes the problem well-posed, even in analysis based on very short time intervals. The sparsity does not compromise the full correlation of audio-visual signals. The algorithm is essentially free of user-defined parameters. The numerical scheme is efficient, based on linear programming. We demonstrate the algorithm in experiments using real data.

This paper is organized as follows. Section II describes CCA, which is a useful tool in multimodal processing. In Section III, we show an alternative yet equivalent formulation of CCA. This formulation serves our analysis, as it highlights the ill-posedness of the problem, revealing the need for regularization. Section IV is dedicated to the exploration of several standard regularization methods. We argue that while such regularization methods lead to unique solutions, the results are far from satisfactory in general. Section V presents the main contribution of this paper. It describes how sparsity of the solution can lead to more effective localization and fully correlated results. In Section VI we extend the analysis to cases where full correlation is not possible. Section VII unveils a fundamental *chorus ambiguity*. Section VIII presents some experimental demonstrations based on real data. We conclude with a brief discussion in Section IX. Partial results appear in [26].

II. Canonical Correlation and Its Limitations

An important tool for exploring the relationship between two modalities is CCA. In this section we describe CCA, and the reason for its importance and popularity in multimodal analysis [1]–[5]. We then expose a fundamental limitation of CCA in the context of our problem. CCA deals with correlation between two random vectors. The vectors can be of different nature and dimensions, such as audio and visual signals. Let $\mathbf{v}$ represent an instantaneous visual signal corresponding to a single frame, e.g., by pixel values or by its wavelet coefficients. Let $\mathbf{a}$ represent a corresponding audio signal, e.g., by the intensity of different audio bands (temporal slices of the periodogram) covering a temporal interval that matches a video frame. Both signals are considered as random vectors, due to their temporal variations.[1] Each of these vectors is projected onto a one dimensional subspace $\mathbf{w}_v$ and $\mathbf{w}_a$, respectively. The result of these projections is a pair of two scalar random variables, $\mathbf{v}^T\mathbf{w}_v$ and $\mathbf{a}^T\mathbf{w}_a$, where T denotes transposition. The normalized correlation coefficient of these two variables defines the canonical correlation [27], [28] between $\mathbf{v}$ and $\mathbf{a}$,

$$\begin{aligned}\rho &\equiv \frac{E\left[\mathbf{w}_v^T\mathbf{v}\mathbf{a}^T\mathbf{w}_a\right]}{\sqrt{E\left[\mathbf{w}_v^T\mathbf{v}\mathbf{v}^T\mathbf{w}_v\right]E\left[\mathbf{w}_a^T\mathbf{a}\mathbf{a}^T\mathbf{w}_a\right]}}\\ &= \frac{\mathbf{w}_v^T\mathbf{C}_{va}\mathbf{w}_a}{\sqrt{\mathbf{w}_v^T\mathbf{C}_{vv}\mathbf{w}_v\mathbf{w}_a^T\mathbf{C}_{aa}\mathbf{w}_a}} \qquad (1)\end{aligned}$$

where E denotes expectation. Here $\mathbf{C}_{vv}$ and $\mathbf{C}_{aa}$ are the covariance matrices of $\mathbf{v}$ and $\mathbf{a}$, respectively, while $\mathbf{C}_{va}$ is the cross-covariance matrix of the vectors.

Maximization of the correlation ρ seeks the subspaces $\mathbf{w}_v$ and $\mathbf{w}_a$ that optimize (1). Note that the solution is scale invariant due to the normalization of (1). This optimization problem has a closed form solution, based on its formulation as an eigenvalues problem [27]

$$\begin{aligned}\mathbf{C}_{vv}^{-1}\mathbf{C}_{va}\mathbf{C}_{aa}^{-1}\mathbf{C}_{av}\mathbf{w}_v &= \rho^2\mathbf{w}_v\\ \mathbf{C}_{aa}^{-1}\mathbf{C}_{av}\mathbf{C}_{vv}^{-1}\mathbf{C}_{va}\mathbf{w}_a &= \rho^2\mathbf{w}_a. \qquad (2)\end{aligned}$$

[1] Each of the vectors $\mathbf{v}$ and $\mathbf{a}$ is assumed to have zero expectation. Numerically, this can be achieved by removal of each vectors' mean prior to application of CCA.

Maximizing $|\rho|$ is equivalent to finding the largest eigenvalue and its corresponding eigenvectors. In the optimal $\mathbf{w}_v$, the components that have the largest magnitude indicate the visual components that best correlate with the projection of $\mathbf{a}$, and vice-versa. Note that a correlation value ρ and its opposite $-\rho$ correspond to the same eigenvalue and eigenvectors in (2). Hence, the range $0 \le \rho \le 1$ is equivalent to $-1 \le \rho \le 0$.

At first sight, CCA may appear to be a good tool for correlating audio and visual signals. The projection of feature vectors can bridge the huge dimensionality gap between sound and pictures. Moreover, CCA amounts to an eigensystem solution. Owing to these attractive characteristics, methods based on projections of feature vectors have been the core of several audio-visual algorithms [8], [11]–[13]. However, CCA and its related methods [13] have a serious shortcoming. The fundamental problem is the *scarcity of data* available in short time intervals, which is often *insufficient* for reliably estimating the statistics of the signals. To see this, note that $\mathbf{C}_{vv}$, $\mathbf{C}_{aa}$ and $\mathbf{C}_{va}$ should be learned from the data. In practice, $\mathbf{C}_{vv}$ is estimated as the empirical matrix

$$\widehat{\mathbf{C}}_{vv} = (1/N_F)\sum_{t=1}^{N_F} \mathbf{v}(t)\mathbf{v}^T(t) \tag{3}$$

where $\mathbf{v}(t)$ is the vector of visual features at time (frame) t and N_F is the total number of frames used for the estimation. For a reliable representation of typical images, at least thousands of visual features are needed. To reliably learn the statistics of $\mathbf{v}$ and get a full rank matrix $\widehat{\mathbf{C}}_{vv}$ to be inverted, as required in (2), we must use at least that number of frames. This imposes minutes-long sequences, while assuming stationarity.

To grasp dynamic events, short time intervals should be used (small N_F), but then we run into a problem of data shortage. The matrix $\widehat{\mathbf{C}}_{vv}$ becomes highly rank deficient, hence (2) cannot be solved, making CCA ill-posed. Technically, the rank deficiency of $\widehat{\mathbf{C}}_{vv}$ can be bypassed by regularization, e.g., by weighted averaging of $\widehat{\mathbf{C}}_{vv}$ with an identity matrix [29]–[31]. Such operations do not overcome the fundamental problem of unreliable statistics. They yield an arbitrary solution that compromises the correlation ρ. As we show in Section IV, such regularization suffers from serious shortcomings, in the context of our problem.

The gap between the amount of data and degrees of freedom is not limited to CCA. It affects optimization of MI just as well. Hence, in some studies, very small images $\mathcal{O}(50 \times 50)$ have been used, out of which only a few dozen features were selected by aggressive pruning or face detection algorithms (the latter limiting audio analysis to speech). In contrast, we seek localization of general sources, while handling intricate details and dynamics.

III. CCA—An Equivalent Formulation

A. The Equivalent Formulation

Before approaching our suggested solution, let us first present an equivalent formulation to CCA that provides more insight. The motivation for this alternative formulation will become evident as we turn to the end of Sections IV and V, to handle the ill-posedness of CCA. Let N_v be the number of visual features. Define the matrix $\mathbf{V} \in \mathcal{R}^{N_F \times N_v}$, where row t contains the vector $\mathbf{v}^T(t)$. Similarly, define $\mathbf{A} \in \mathcal{R}^{N_F \times N_a}$, where row t contains the coefficients of the audio signal $\mathbf{a}^T(t)$, and N_a is the number of audio features. Note that $\mathbf{v}(t)$ and $\mathbf{a}(t)$ are time series, so matrices $\mathbf{V}$ and $\mathbf{A}$ contain each time point in their rows. Defining the empirical covariances matrices $\hat{\mathbf{C}}_{vv} = \mathbf{V}^T\mathbf{V}$, $\hat{\mathbf{C}}_{aa} = \mathbf{A}^T\mathbf{A}$ and $\hat{\mathbf{C}}_{va} = \hat{\mathbf{C}}_{av}^T = \mathbf{V}^T\mathbf{A}$, the empirical canonical correlation[2] (1) becomes

$$\hat{\rho} = \frac{\mathbf{w}_v^T(\mathbf{V}^T\mathbf{A})\mathbf{w}_a}{\sqrt{\mathbf{w}_v^T(\mathbf{V}^T\mathbf{V})\mathbf{w}_v\mathbf{w}_a^T(\mathbf{A}^T\mathbf{A})\mathbf{w}_a}}. \tag{4}$$

CCA seeks to maximize $|\hat{\rho}|$. As we show next, maximizing $|\hat{\rho}|$ is equivalent to minimizing the penalty function

$$G(\mathbf{w}_v, \mathbf{w}_a) = \frac{\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\mathbf{w}_a\|_2^2}{\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\mathbf{w}_a\|_2^2} \tag{5}$$

with respect to $\mathbf{w}_v$ and $\mathbf{w}_a$, where $\|\cdot\|_2$ is the ℓ^2-norm.[3] To prove this, we null the derivatives of $G(\mathbf{w}_v, \mathbf{w}_a)$

$$\frac{\partial}{\partial \mathbf{w}_v} G(\mathbf{w}_v, \mathbf{w}_a) = 0, \quad \frac{\partial}{\partial \mathbf{w}_a} G(\mathbf{w}_v, \mathbf{w}_a) = 0. \tag{6}$$

This leads to

$$\begin{aligned} &\mathbf{V}^T(\mathbf{V}\mathbf{w}_v - \mathbf{A}\mathbf{w}_a)\left(\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\mathbf{w}_a\|_2^2\right) \\ &\quad - \mathbf{V}^T\mathbf{V}\mathbf{w}_v\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\mathbf{w}_a\|_2^2 = 0 \qquad (7) \\ &- \mathbf{A}^T(\mathbf{V}\mathbf{w}_v - \mathbf{A}\mathbf{w}_a)\left(\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\mathbf{w}_a\|_2^2\right) \\ &\quad - \mathbf{A}^T\mathbf{A}\mathbf{w}_a\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\mathbf{w}_a\|_2^2 = 0 \qquad (8) \end{aligned}$$

hence

$$\mathbf{V}^T\mathbf{V}\mathbf{w}_v - \mathbf{V}^T\mathbf{A}\mathbf{w}_a = \frac{\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\mathbf{w}_a\|_2^2}{\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\mathbf{w}_a\|_2^2}\mathbf{V}^T\mathbf{V}\mathbf{w}_v \tag{9}$$

$$-\mathbf{A}^T\mathbf{V}\mathbf{w}_v + \mathbf{A}^T\mathbf{A}\mathbf{w}_a = \frac{\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\mathbf{w}_a\|_2^2}{\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\mathbf{w}_a\|_2^2}\mathbf{A}^T\mathbf{A}\mathbf{w}_a. \tag{10}$$

Using the empirical covariance matrices and the definition of $G(\mathbf{w}_v, \mathbf{w}_a)$, we obtain

$$\hat{\mathbf{C}}_{vv}\mathbf{w}_v - \hat{\mathbf{C}}_{va}\mathbf{w}_a = G\hat{\mathbf{C}}_{vv}\mathbf{w}_v \tag{11}$$

implying

$$\mathbf{w}_v = \frac{1}{1-G}\hat{\mathbf{C}}_{vv}^{-1}\hat{\mathbf{C}}_{va}\mathbf{w}_a. \tag{12}$$

An analogous derivation for (10) yields

$$\mathbf{w}_a = \frac{1}{1-G}\hat{\mathbf{C}}_{aa}^{-1}\hat{\mathbf{C}}_{av}\mathbf{w}_v. \tag{13}$$

Equations (12) and (13) yield

$$\begin{aligned} \hat{\mathbf{C}}_{vv}^{-1}\hat{\mathbf{C}}_{va}\hat{\mathbf{C}}_{aa}^{-1}\hat{\mathbf{C}}_{av}\mathbf{w}_v &= (1-G)^2\mathbf{w}_v \\ \hat{\mathbf{C}}_{aa}^{-1}\hat{\mathbf{C}}_{av}\hat{\mathbf{C}}_{vv}^{-1}\hat{\mathbf{C}}_{va}\mathbf{w}_a &= (1-G)^2\mathbf{w}_a. \end{aligned} \tag{14}$$

Note that (14) is equivalent to the CCA set of equations (2), with $\rho^2 = (1-G)^2$. Thus, an extremum of G is *equivalent* to

[2] Strictly speaking, the definition for $\widehat{\mathbf{C}}_{vv}$, $\widehat{\mathbf{C}}_{aa}$ and $\widehat{\mathbf{C}}_{va}$ should be normalized by N_F. However, this constant is factored out in (4), and is thus discarded throughout the paper.

[3] Note that $0 \le G(\mathbf{w}_v, \mathbf{w}_a) \le 2$. The proof is given in Section A of the Appendix.

an extremum of ρ. Moreover, finding the maximum correlation (i.e., the largest eigenvalue ρ^2) is equivalent to minimizing[4] G. It can be shown that the range of $0 \leq G \leq 1$ is equivalent to the range $0 \leq \rho \leq 1$, while $1 \leq G \leq 2$ is equivalent to $-1 \leq \rho \leq 0$. As we discussed in Section II, these two ranges are equivalent. Thus, the solution that maximizes G in the domain $1 \leq G \leq 2$ is equivalent to the one minimizing G when $0 \leq G \leq 1$. Hence, in this paper we can focus on minimizing G towards zero.

To gain intuition into the equivalence of (4) and (5), note that (5) is minimized if the projected video $\mathbf{V}\mathbf{w}_v$ is as close as possible to the projected audio $\mathbf{A}\mathbf{w}_a$, in the ℓ^2 sense. Hence, we seek linear dependency between $\mathbf{V}\mathbf{w}_v$ and $\mathbf{A}\mathbf{w}_a$, as expected in high correlation. The denominator in (5) serves to avoid trivial solutions, and to properly use the energies of the two projections. This is analogous to the correlation normalization in (1).

Before proceeding, we note that there is an alternative formulation to CCA, called *principal angles* [32]–[34]. For the principal angles approach, an alternative formulation was proposed in [34], which is the constrained optimization

$$\max_{\mathbf{w}_a,\mathbf{w}_v} \left\{\mathbf{w}_v^T\mathbf{V}^T\mathbf{A}\mathbf{w}_a\right\} \text{ subject to } \|\mathbf{V}\mathbf{w}_v\|_2^2=1,\ \|\mathbf{A}\mathbf{w}_a\|_2^2=1. \tag{15}$$

We prefer working with the unconstrained optimization of (5), rather than (15), since the former is exactly equivalent to CCA in its classical form (2).

B. The Ill-Posedness of CCA

CCA has limitations when working with a rank deficient matrix $\hat{\mathbf{C}}_{vv}$. In the wider context of cross modal analysis, this occurs, for example, if too few temporal samples of meteorological events are used, or if details of just a few subjects are known in a multi-parameter medical study. In the context of audio-visual correlation, this occurs when short time intervals are used. Here, the number of representation features (at each frame) is expected to be much larger than the number of frames in the time interval $(N_F \ll N_v)$. Let us analyze this ill-posedness using (5). We first focus on the cases where $N_a = 1$, i.e., the audio is characterized by a single feature. The case of multiple audio bands $(N_a > 1)$ is treated in Section V-B.

When $N_a = 1$ we may set $\mathbf{w}_a = 1$ (where $\mathbf{w}_a$ is a scalar), since the penalty function in (5) is scale invariant (multiplying $\mathbf{w}_v$ and $\mathbf{w}_a$ by the same constant does not change the function's value). Thus, (5) becomes

$$G(\mathbf{w}_v) = \frac{\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\|_2^2}{\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\|_2^2}. \tag{16}$$

The denominator of (16) is necessarily not zero. The reason is that $\|\mathbf{A}\| \neq 0$, otherwise audio does not exist and cross-modal analysis is not possible. Define the numerator

$$g(\mathbf{w}_v) = \|\mathbf{V}\mathbf{w}_v - \mathbf{A}\|_2^2. \tag{17}$$

Suppose that a vector $\mathbf{w}_v$ exists[5] such that $g(\mathbf{w}_v) = 0$. This vector yields $G(\mathbf{w}_v) = 0$, since the denominator of (16) is necessarily nonzero. Hence, this solution yields complete coherence, $|\hat{\rho}| = 1$, as desired. Requiring $g(\mathbf{w}_v) = 0$ implies

$$\mathbf{V}\mathbf{w}_v = \mathbf{A}. \tag{18}$$

Since $N_a = 1$, $\mathbf{A}$ is a *column* vector of length N_F. As discussed in Section II, $N_v \gg N_F$, where N_v is the length of $\mathbf{w}_v$. Therefore, in the set of linear equations (18), the number of equations is much smaller than the number of unknowns, yielding an underdetermined linear set of equations. If $\mathbf{V}$ is full rank, the number of possible solutions is infinite. To conclude: due to the scarce data, there are infinite number of combinations of visual features that appear to completely correlate with the audio!

How probable is the scenario of having $g(\mathbf{w}_v) = 0$? For $N_v \gg N_F$, most chances are that $\text{rank}(\mathbf{V}) = N_F$, since $\mathbf{V}$ is stochastic due to scene dynamics. This generally guarantees that $\mathbf{A}$ is in the span of the $\mathbf{V}$ column space. Thus, it is highly probable that $g(\mathbf{w}_v)$ has a zero. In fact, noise in the visual data guarantees this outcome, as it causes the rank to become full. However, visual noise implies strong correlation of "junk" features to the audio.[6]

IV. Attempting Standard Regularizations

Since CCA of scarce data is ill-posed, regularization should be imposed. Regularization has the role of choosing the best vector among the infinite space of potential solutions, according to some criterion. Next, several types of standard regularization techniques are discussed, as well as their drawbacks. Our alternative approach, which is stronger in the context of localization is introduced in Section V.

A. Minimum Energy Regularization Using an ℓ^2 Term

A common regularization of underdetermined problems is to prefer the minimal energy solution [33], [35]. In our case, this would be

$$\min \|\mathbf{w}_v\|_2 \quad \text{subject to} \quad \mathbf{V}\mathbf{w}_v = \mathbf{A}. \tag{19}$$

The constraint $\mathbf{V}\mathbf{w}_v = \mathbf{A}$ nulls the numerator of (16), thus leading to a solution having full correlation. The ℓ^2 term in (19) is the imposed regularization. The $\mathbf{w}_v$ that solves (19) is well known in the literature and may be found using one of several possible techniques, such as the Moore-Penrose pseudoinverse, SVD, or QR factorization [33].

In the context of the audio-visual problem, (19) results in *poor visual localization*. The reason is that the ℓ^2 criterion seeks to spread the energy of $\mathbf{w}_v$ over many small-valued visual components, rather than concentrating energy on a few dominant ones. To obtain some intuition, this phenomenon is depicted in the left part of Fig. 2 for $N_v = 2$ and $N_F = 1$. In this figure, a straight line describes the linear constraint $\mathbf{V}\mathbf{w}_v = \mathbf{A}$. The minimum of the ℓ^2-norm is obtained in point B, which has substantial energy in all components. This nature is contrary to common audio-visual scenarios, where visual events associated with sound are often very *local*. They typically reside in small areas (few components) of the frame. Indeed, the inadequacy of

[4] Note that G is real and nonnegative, by definition.

[5] The complementary cases are treated in Section VI.

[6] On the extreme, if $\mathbf{V}$ is just a noise matrix, it has full rank, nulling $g(\mathbf{w}_v)$, but yielding meaningless results.

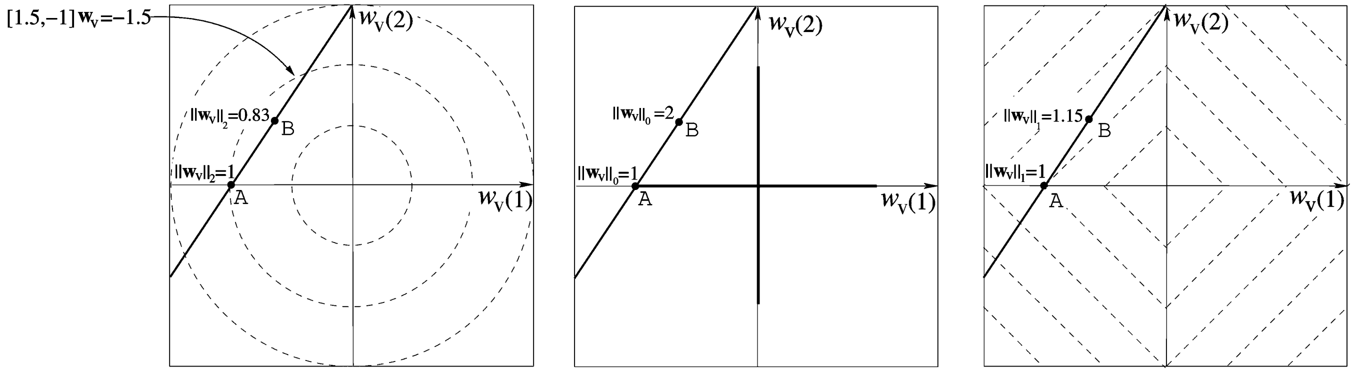


Fig. 2. A 2-D example of optimization under [left] ℓ^2-norm [middle] ℓ^0-norm [right] ℓ^1-norm. The dashed contours represent iso-norm levels. On the linear constraint $\mathbf{V}\mathbf{w}_v = \mathbf{A}$ (solid line), point B minimizes $\|\mathbf{w}_v\|_2$, but it has substantial energy in all components. In contrast, point A on the solid line is the sparsest (minimum $\|\mathbf{w}_v\|_0$), and also satisfied minimum $\|\mathbf{w}_v\|_1$. The ℓ^1 criterion is convex.

this criterion is further demonstrated in the experiments detailed in Section VIII.

B. Regularization Using the Identity Matrix

As described in (2), CCA requires the inversion[7] of $\hat{\mathbf{C}}_{vv}$. This matrix is singular and highly rank-defficient. One way to overcome this problem is to regularize $\hat{\mathbf{C}}_{vv}$ by defining an invertible version

$$\widetilde{\mathbf{C}}_{vv} = \hat{\mathbf{C}}_{vv} + \epsilon\mathbf{I} \tag{20}$$

where $\mathbf{I}$ is the identity matrix and ϵ is an arbitrary small number. This has been a common approach in CCA regularization [29]–[31]. It makes the covariance matrix full rank and invertible.[8] However, as we show in this paper, this regularization reduces the correlation value (destroying the complete coherence) and has a resemblance to the ℓ^2 regularization posed earlier in (19).

To assess this regularization, we relate it to the penalty function in (5). Recalling that $\hat{\mathbf{C}}_{vv} = \mathbf{V}^T\mathbf{V}$, we can obtain (20) by defining a matrix $\widetilde{\mathbf{V}}$ of size $N_v \times N_v$, having the form

$$\widetilde{\mathbf{V}} = \begin{bmatrix} \mathbf{V}_{N_F \times N_v} \\ \sqrt{\epsilon}\mathbf{I}_{(N_v - N_F)\times N_v} \end{bmatrix} \tag{21}$$

and then $\widetilde{\mathbf{C}}_{vv} = \widetilde{\mathbf{V}}^T\widetilde{\mathbf{V}}$. Suppose we use $\widetilde{\mathbf{C}}_{vv}$ and $\widetilde{\mathbf{V}}$ defined by (20) and (21) instead of the original matrices $\mathbf{V}$ and $\hat{\mathbf{C}}_{vv}$. Insert matrix $\widetilde{\mathbf{V}}$ into (16), and define $\widetilde{\mathbf{A}}$ to be a zero-padded version of $\mathbf{A}$. This leads to a *regularized cost function*

$$\widetilde{G}(\mathbf{w}_v) = \frac{\|\widetilde{\mathbf{V}}\mathbf{w}_v - \widetilde{\mathbf{A}}\|_2^2}{\|\widetilde{\mathbf{V}}\mathbf{w}_v\|_2^2 + \|\widetilde{\mathbf{A}}\|_2^2} = \frac{\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\|_2^2 + \epsilon\|\mathbf{w}_v\|_2^2}{\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\|_2^2 + \epsilon\|\mathbf{w}_v\|_2^2} \tag{22}$$

where

$$\widetilde{\mathbf{A}} = \begin{bmatrix} \mathbf{A} \\ \mathbf{0}_{N_v - N_F} \end{bmatrix}. \tag{23}$$

The number ϵ is small, while the audio data $\mathbf{A}$ is assumed to contain significant energy. We thus assume that $\epsilon\|\mathbf{w}_v\|_2^2 \ll \|\mathbf{A}\|_2^2$. Thus, $\epsilon\|\mathbf{w}_v\|_2^2$ can be omitted from the denominator of (22). However, this term cannot be neglected in the numerator, since $\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\|_2^2$ is a small number (as we are close to full correlation). The regularized cost function becomes

$$\begin{aligned}\widetilde{G}(\mathbf{w}_v) &\approx \frac{\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\|_2^2 + \epsilon\|\mathbf{w}_v\|_2^2}{\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\|_2^2} \\ &= G(\mathbf{w}_v) + \epsilon\frac{\|\mathbf{w}_v\|_2^2}{\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\|_2^2}\end{aligned} \tag{24}$$

where $G(\mathbf{w}_v)$ is the penalty function of the nonregularized matrices. Recall that maximizing the correlation is equivalent to minimizing G, rather than $\widetilde{G}$. On the other hand, minimizing (24) tends to minimize $\|\mathbf{w}_v\|_2^2$ as well. Thus, this method has a strong resemblance to the ℓ^2 regularization given in (19). It may, thus, be prone to a similar energy spread drawback. Moreover, it generally leads to a reduced correlation, as proved in Section B of the Appendix.

V. Sparsity as a Key

"Out of clutter, find simplicity.
From discord, find harmony."
Albert Einstein

As we have shown in the previous section, solving the audio-video correlation problem using the traditional ℓ^2-norm solution, leads to poorly localized results. We now describe our approach, which leads to a unique solution based on a spatial sparsity criterion. First, we look at cases where $N_a = 1$, i.e., the audio is characterized by a single feature. In Section V-B we extend the analysis to multiple audio bands.

A. A Single Audio Band

When using a single audio band, our goal is to minimize (16). We first discuss cases where the minimum of this function is zero. The case of a non-zero cost function value is discussed in Section VI. As discussed in Section IV-A, (16) has infinitely many possible solutions, all of which have the same correlation

[7] Inversion of $\mathbf{C}_{aa}$ is not a problem in our audio-visual localization scenario. The reason is that the number of audio features is comparable to the number of temporal samples, i.e., $N_a \sim N_F$.

[8] The covariance matrix in its new formulation (20) can be inverted efficiently using the Sherman-Morrison formula [33]. However, it still involves a huge eigenproblem.

value. To overcome this ambiguity, we express *locality* as a requirement that the sought solution is *sparse*,[9] meaning that only a small number of visual features are associated with the audio. Thus, out of the entire space of possible correlated projections, we may aim to solve

$$\min \|\mathbf{w}_v\|_0 \quad \text{subject to} \quad \mathbf{V}\mathbf{w}_v = \mathbf{A} \tag{25}$$

where $\|\cdot\|_0$ is the ℓ^0-norm of a vector space (the number of non-zero vector coefficients). In the simple example depicted on the middle of Fig. 2, the optimal solution according to this criterion (point A) has a single component.[10] Unfortunately, this criterion is not convex, and the complexity of its optimization is exponential [36], [37] in N_v. We bypass this difficulty by turning the problem into a convex one and solving

$$\min \|\mathbf{w}_v\|_1 \quad \text{subject to} \quad \mathbf{V}\mathbf{w}_v = \mathbf{A} \tag{26}$$

where ℓ^1 is used instead of ℓ^0. This idea is known by the name *basis pursuit* [40]. In the right part of Fig. 2, the solution optimizing this alternative criterion has a single component (point A), just as with the ℓ^0 criterion. All other points in the linear constraint $\mathbf{V}\mathbf{w}_v = \mathbf{A}$ have a larger ℓ^1-norm. Thus, there is an apparent equivalence between ℓ^0 and ℓ^1, since both lead to the same optimal vector. Moreover, this figure illustrates the convexity of the ℓ^1 criterion.

In general, the equivalence of the ℓ^0 and ℓ^1 problems (25), (26) has been studied in depth during the last couple of years from a pure mathematical perspective. First contributions in this direction considered deterministic sufficient conditions for this equivalence [36], [37], [41]. More recently, a probabilistic approach has been introduced, showing that equivalence holds true far beyond the limits determined by these sufficient conditions [42], [43]. Further details about the equivalence and its conditions are given in Section C of the Appendix. Owing to this theoretical progress, formulating sparsity using the ℓ^1-norm is effective. We note that there are other approaches for efficiently minimizing (25). For example, the FOCUSS method [44] replaces the ℓ^0-norm by an ℓ^p with $p < 1$. Such methods may be used as well to solve our problem. Anyway, to benefit from convexity we used basis pursuit.

Equation (26) can be given a statistical interpretation, according to which the unknown $\mathbf{w}_v$ is a random vector. Among all possible solutions satisfying $\mathbf{V}\mathbf{w}_v = \mathbf{A}$, one may seek the solution with the highest probability. In this line of thought, each element in $\mathbf{w}_v$ is assumed to be a Laplacian random variable, and $\mathbf{w}_v$ is a combination of N_v independent and identically distributed (i.i.d) such random variables. Such a model is commonly practiced in image processing [40]. However, in this paper we do not follow this interpretation, and rather use a deterministic point of view.

The newly defined formulation (26) can be posed as a *linear programming* problem, and thus can be solved *efficiently*, even for $N_v \gg 1$. This formulation influences the solution energy to concentrate on few visual features which strongly correlate with the audio. It penalizes for dispersed components, particulary the random "junk" features described earlier, e.g., image noise. Moreover, the solution is *unique*, thanks to the convexity of the ℓ^1-norm, except for special cases discussed in Section VII.

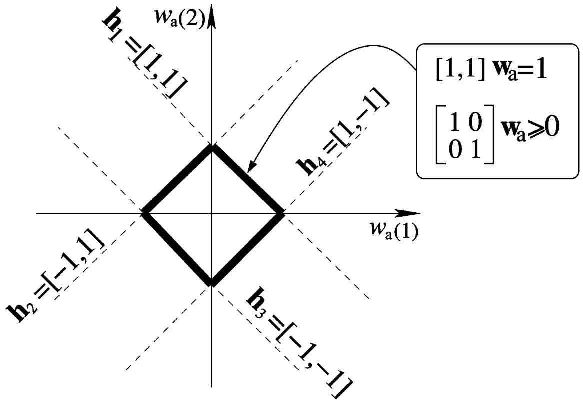


Fig. 3. A 2–D illustration of the faces of the ℓ^1-ball in the audio space.

B. Multiple Audio Bands

We now generalize the analysis of Section V-A to audio signals that are divided into multiple bands. We postpone to Section VI the analysis of scenarios in which the optimal value of the cost function G is nonzero. Here, we analyze cases where the cost function can become zero. This allows us to concentrate on the numerator of (5). The numerator is zero if and only if

$$\mathbf{V}\mathbf{w}_v = \mathbf{A}\mathbf{w}_a. \tag{27}$$

As before, if $\operatorname{rank}(\mathbf{V}) = N_F$, a zero solution of G is guaranteed. As we have claimed in Section V-A, this is a highly probable event. In the unlikely event that no intersection exists between the subspace spanned by the columns of $\mathbf{V}$ and the subspace spanned by $\mathbf{A}$, the cost function G cannot be nulled (see Section VI).

As aforementioned, (27) is prone to a scale ambiguity. To overcome this problem and avoid the trivial solution $\mathbf{w}_a = 0$, we use normalization. A way to achieve this is to limit the search to the ℓ^1-ball of audio-weights, $\|\mathbf{w}_a\|_1 = 1$. The set $\|\mathbf{w}_a\| = 1$ is not convex. To keep enjoying the benefits of convexity in our problem formulation, we break the problem into 2^{N_a} separate ones, where each handles a single face of the audio ℓ^1-ball and is thus convex. As depicted in Fig. 3, the optimization over each face $q \in [1, 2^{N_a}]$ can be posed as

$$s_q = \min \|\mathbf{w}_v\|_1 \quad \text{subject to} \\ \left\{\mathbf{V}\mathbf{w}_v = \mathbf{A}\mathbf{w}_a, \mathbf{h}_q^T\mathbf{w}_a = 1, \mathbf{H}_q\mathbf{w}_a \geq 0\right\} \tag{28}$$

where $\mathbf{h}_q$ is a vector of length N_a, and $\mathbf{H}_q$ is a diagonal matrix whose diagonal is $\mathbf{h}_q$. The vector set $\{\mathbf{h}_q\}_{q=1}^{2^{N_a}}$ comprises the 2^{N_a} different combinations of the N_a-tuples binary sequences with ±1 as their entries. Since all the constraints are linear, (28) is solved for each q using linear programming.

Recall that for our audio-visual localization method, we should optimize the visual sparsity over the audio ℓ^1-ball. This

[9] In video analysis, sparsity is enhanced using a wavelet representation of temporal-difference images.

[10] This example should be viewed with caution: there are two intersections of the linear constraint with the axes of $\mathbf{w}_v$, both considered as global minimizers of ℓ^0. This lack of uniqueness is due to the low dimension (2–D of this example). As the dimensions of the problem grow, uniqueness becomes possible. For details see [36]–[39].

is done by running (28) over all[11] values of q, and then selecting the optimal q by

$$\hat{q} = \arg\min s_q. \tag{29}$$

The unique vectors $\mathbf{w}_v$ and $\mathbf{w}_a$ that we seek are then derived by using this specific $\hat{q}$ in (28). We stress that our goal is to localize *visual* events (based on audio cues), while processing of audio is of secondary importance here. This distinction enables us to use a coarse representation of the audio. Hence, only a small number of audio bands N_a is required. For this reason, the computations are tolerable despite the $\mathcal{O}(2^{N_a})$ complexity.

VI. A Non-Zero Cost Function Value

So far we considered solutions $\mathbf{w}_v$ that null $g(\mathbf{w}_v)$. This nulling is very likely, by a full rank of $\mathbf{V}$, as explained in Section III. However, there is a chance, even if it is low, that $\mathbf{V}$ is not full rank, hence, no solution is fully correlated. For the sake of completeness, we show that this case can be handled well by our approach.

A. A Single Audio Band

It follows from Section III that $\min[g(\mathbf{w}_v)] \neq 0$ only if $\mathrm{rank}(\mathbf{V}) < N_F$, and if $\mathbf{A}$ is not in the column span of $\mathbf{V}$. In such cases, we can decompose $\mathbf{A}$ as $\mathbf{A} = \mathbf{A}_{||} + \mathbf{A}_{\perp}$. Here $\mathbf{A}_{||}$ is in the subspace spanned by the columns of $\mathbf{V}$, while $\mathbf{A}_{\perp}$ is orthogonal to $\mathbf{V}$. Thus,

$$g(\mathbf{w}_v) = ||\mathbf{V}\mathbf{w}_v - \mathbf{A}||_2^2 = ||\mathbf{V}\mathbf{w}_v - \mathbf{A}_{||}||_2^2 + ||\mathbf{A}_{\perp}||_2^2 \tag{30}$$

and (16) becomes

$$G(\mathbf{w}_v) = \frac{||\mathbf{V}\mathbf{w}_v - \mathbf{A}||_2^2}{||\mathbf{V}\mathbf{w}_v||_2^2 + ||\mathbf{A}||_2^2} = \frac{||\mathbf{V}\mathbf{w}_v - \mathbf{A}_{||}||_2^2 + ||\mathbf{A}_{\perp}||_2^2}{||\mathbf{V}\mathbf{w}_v||_2^2 + ||\mathbf{A}_{||}||_2^2 + ||\mathbf{A}_{\perp}||_2^2}. \tag{31}$$

Note that the audio component $\mathbf{A}_{\perp}$ does not correlate with any of the visual features. As such, it can be discarded as irrelevant. The remaining signal $\mathbf{A}_{||}$ is a projected version of the original audio for which the solution to $\mathbf{V}\mathbf{w}_v = \mathbf{A}_{||}$ exists. Thus, $\mathbf{A}$ is essentially *projected* to the column space of $\mathbf{V}$, as a "denoising" preprocess. This suggests that we handle the rank-deficient $\mathbf{V}$ matrix case by such a projection, and then proceed as in (26) where we use $\mathbf{A}_{||}$ instead of $\mathbf{A}$.

As we show now, this line of reasoning is in fact optimal up to a scale. We are interested in characterizing the set of minimizers $\mathbf{w}_v$ of (31). Recall that $\mathbf{A}_{\perp}$ is not spanned by the columns of $\mathbf{V}$. Thus, no matter what $\mathbf{w}_v$ is, the term $\mathbf{V}\mathbf{w}_v$ is necessarily orthogonal to $\mathbf{A}_{\perp}$. In general, the solution $\mathbf{w}_v$ satisfies the relation $\mathbf{V}\mathbf{w}_v = \alpha\mathbf{A}_{||} + \mathbf{Z}$, where α is a scalar, and $\mathbf{Z}$ is an arbitrary vector perpendicular to both $\mathbf{A}_{||}$ and $\mathbf{A}_{\perp}$. Thus, $G(\mathbf{w}_v)$ in (31) becomes the function $G(\alpha, \mathbf{Z})$

$$G(\alpha, \mathbf{Z}) = \frac{||\alpha\mathbf{A}_{||} + \mathbf{Z} - \mathbf{A}_{||}||_2^2 + ||\mathbf{A}_{\perp}||_2^2}{||\alpha\mathbf{A}_{||} + \mathbf{Z}||_2^2 + ||\mathbf{A}_{||}||_2^2 + ||\mathbf{A}_{\perp}||_2^2}. \tag{32}$$

We need to find α and $\mathbf{Z}$ that minimize this function. We thus derive equations that null the partial derivatives of $G(\alpha, \mathbf{Z})$ with respect to α and $\mathbf{Z}$. Handling $\mathbf{Z}$ first, we rearrange G

$$\begin{aligned} G(\alpha, \mathbf{Z}) &= \frac{(\alpha - 1)^2||\mathbf{A}_{||}||_2^2 + ||\mathbf{Z}||_2^2 + ||\mathbf{A}_{\perp}||_2^2}{(1 + \alpha^2)||\mathbf{A}_{||}||_2^2 + ||\mathbf{Z}||_2^2 + ||\mathbf{A}_{\perp}||_2^2} \\ &= \frac{(\alpha - 1)^2 + \frac{||\mathbf{Z}||_2^2}{||\mathbf{A}_{||}||_2^2} + \frac{||\mathbf{A}_{\perp}||_2^2}{||\mathbf{A}_{||}||_2^2}}{(1 + \alpha^2) + \frac{||\mathbf{Z}||_2^2}{||\mathbf{A}_{||}||_2^2} + \frac{||\mathbf{A}_{\perp}||_2^2}{||\mathbf{A}_{||}||_2^2}}. \end{aligned} \tag{33}$$

Here we have exploited the fact that the ℓ^2-norm is separable when dealing with two orthogonal vectors ($\mathbf{A}_{||}$ and $\mathbf{Z}$ in this case). To simplify this expression, let us define $r \equiv ||\mathbf{A}_{\perp}||_2^2/||\mathbf{A}_{||}||_2^2$ and $k(\mathbf{Z}) \equiv ||\mathbf{Z}||_2^2/||\mathbf{A}_{||}||_2^2$, hence

$$G(\alpha, \mathbf{Z}) = \frac{(\alpha - 1)^2 + k(\mathbf{Z}) + r}{(1 + \alpha^2) + k(\mathbf{Z}) + r}. \tag{34}$$

It follows that

$$\frac{\partial G(\alpha, \mathbf{Z})}{\partial \mathbf{Z}} = \frac{2\mathbf{Z}}{||\mathbf{A}_{||}||_2^2} \cdot \frac{2\alpha}{[(1 + \alpha^2) + k(\mathbf{Z}) + r]^2} \tag{35}$$

where we used $\partial k(\mathbf{Z})/\partial \mathbf{Z} = 2\mathbf{Z}/||\mathbf{A}_{||}||_2^2$. We seek optimization of G. We, thus, require nulling of (35). Hence $\mathbf{Z} = 0$. Furthermore, handling α yields

$$\frac{\partial G(\alpha)}{\partial \alpha} = \frac{2(\alpha^2 - 1 - k(\mathbf{Z}) - r)}{[(1 + \alpha^2) + k(\mathbf{Z}) + r]^2}. \tag{36}$$

Hence

$$\frac{\partial G(\alpha)}{\partial \alpha} = 0 \Rightarrow \alpha_{\mathrm{opt}} = \sqrt{1 + k(\mathbf{Z}) + r} \tag{37}$$

i.e.,

$$\alpha_{\mathrm{opt}} = \sqrt{1 + \frac{||\mathbf{Z}||_2^2}{||\mathbf{A}_{||}||_2^2} + \frac{||\mathbf{A}_{\perp}||_2^2}{||\mathbf{A}_{||}||_2^2}}. \tag{38}$$

Since $\mathbf{Z} = 0$, then $\alpha_{\mathrm{opt}} = \sqrt{1 + ||\mathbf{A}_{\perp}||_2^2/||\mathbf{A}_{||}||_2^2}$. To recap, these values of $\mathbf{Z}$ and α result from minimization of (31), when $\mathbf{V}\mathbf{w}_v = \alpha\mathbf{A}_{||} + \mathbf{Z}$. This means that the correlated audio-visual features satisfy

$$\mathbf{V}\mathbf{w}_v = \alpha\mathbf{A}_{||} = \left(\sqrt{1 + ||\mathbf{A}_{\perp}||_2^2/||\mathbf{A}_{||}||_2^2}\right)\mathbf{A}_{||}. \tag{39}$$

To conclude, if G cannot be nulled, then the set of minimizers $\mathbf{w}_v$ of (31) is given by (39). Since (39) minimizes G, it maximizes the correlation. The scalar α does not influence the localization result, but only the overall scale of $\mathbf{w}_v$. Thus, the results obtained using our algorithm are consistent, up to a scale.

B. Multiple Audio Bands

In the multiple audio band problem, the vector $\mathbf{w}_a$ is unknown. However, from the discussion leading to (39), the optimal $\mathbf{w}_a$ should make the projected audio parallel to the projected visual signal. Thus, we force this parallelism by con-

[11] Actually, there is no need to scan all 2^{N_a} values of q. Due to the scale ambiguity mentioned above, $\mathbf{h}_q$ and $-\mathbf{h}_q$ yield the same results. Hence it is sufficient to scan $2^{N_a - 1}$ nonequivalent values of q.

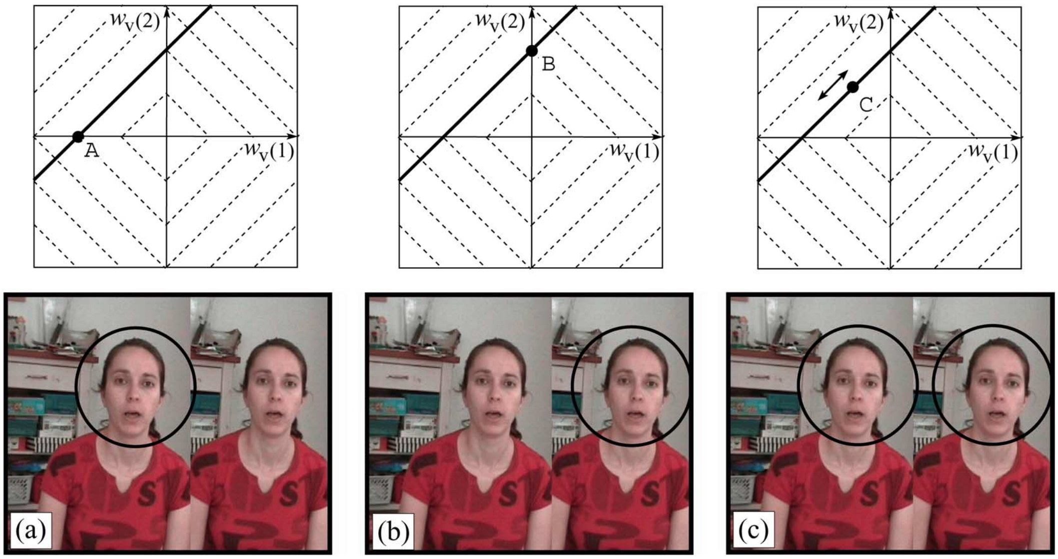


Fig. 4. The chorus ambiguity under the ℓ^1-norm. The top row is a two-pixels scene and the bottom shows a human chorus. (a) Detecting the left person. (b) Detecting the right person. (c) Detecting both.

straining (28), and adapt the formulation to the case where the penalty function is non-zero.

Let the space spanned by the columns of $\mathbf{A}$ be denoted by $\mathcal{A}$. Decompose this space into two orthogonal subspaces $\mathcal{A}_{||}$ and $\mathcal{A}_{\perp}$, where $\mathcal{A}_{||}$ spans the projected audio subspace $\mathbf{A}\mathbf{w}_a$. Define $\mathbf{A}_{||}$ and $\mathbf{A}_{\perp}$ as matrices whose columns span $\mathcal{A}_{||}$ and $\mathcal{A}_{\perp}$, respectively. Similarly to (39), parallelism means that

$$\mathbf{V}\mathbf{w}_v = \beta \mathbf{A}_{||}\mathbf{w}_a \tag{40}$$

where β is a scalar. Thus, the inner product between $\mathbf{V}\mathbf{w}_v$ and the orthogonal audio space (spanned by $\mathbf{A}_{\perp}$) must be zero

$$\mathbf{A}_{\perp} \cdot \mathbf{V}\mathbf{w}_v = 0. \tag{41}$$

We use (40) and (41) as new constraints. Combining these constraints, (28) becomes

$$s_q = \min \ ||\mathbf{w}_v||_1 \quad \text{subject to}$$
$$\left\{\mathbf{V}\mathbf{w}_v = \mathbf{A}_{||}\mathbf{w}_a, \mathbf{A}_{\perp}\cdot\mathbf{V}\mathbf{w}_v = 0, \mathbf{h}_q^T\mathbf{w}_a = 1, \mathbf{H}_q\mathbf{w}_a \geq 0\right\}. \tag{42}$$

Hence, the same algorithm introduced earlier can be used in the general case discussed here.

VII. The Chorus Ambiguity

Consider a chorus of identical people singing in synchrony the same song. In this case, the audio track corresponds well to several spatially distinct clusters of pixels (faces of the chorus members). Which pixels would you choose as the ones achieving successful localization? This scenario poses a fundamental ambiguity for any localization algorithm: the result could pinpoint any single person or several of them. In this special scenario all these results are equally acceptable. We term this phenomenon as the *chorus ambiguity*, and it stands for the loss of the localization uniqueness. Such scenario can also occur in events which are not audio-visual.

Our algorithm as posed in (26), (28), and (29) has this characteristic, just as well. Referring to Fig. 2, this case occurs when the linear constraint $\mathbf{V}\mathbf{w}_v = \mathbf{A}$ aligns with a face of a visual ℓ^1 ball. Mathematically, the implication is that for this special scenario, the problem in (26) does not have a unique solution, but rather a set of them. This case is demonstrated in Fig. 4 for a two-pixels scene (top row) and for a chorus of two people (bottom row). In this illustration, three solution types in the two-pixels scene are represented, denoted by A, B and C. Types A and B represent exclusive detection of only a single pixel, while type C represents all solutions that are a convex superposition of A and B. Analogously, in the two people chorus, types A and B represent an exclusive detection of a single person, while type C represents detection of the entire chorus (with some weight ratio between members).

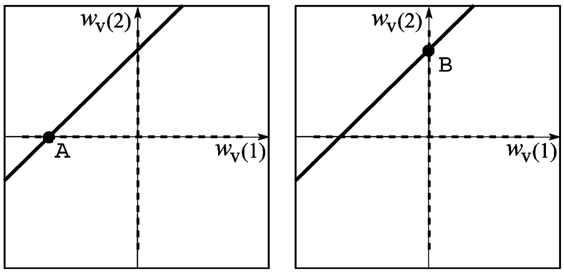


Fig. 5. The chorus ambiguity under the ℓ^0-norm. There are only exclusive detections, which correspond to points A and B in Fig. 4.

We can see that the problem of (25) has only one type of solution, as demonstrated in Fig. 5—that of exclusive detection. In the general chorus case, the ℓ^0 criterion can lock into any single person in the chorus, while the ℓ^1 result can spread the detections between several of them. Thus, in this case the equivalence between ℓ^1 and ℓ^0 breaks down. A mathematical insight to this phenomenon can be found in [36], [37]. Still, this effect does not hinder the optimization process that we have posed: the linear programming converges to one of these solutions, depending on the initialization.

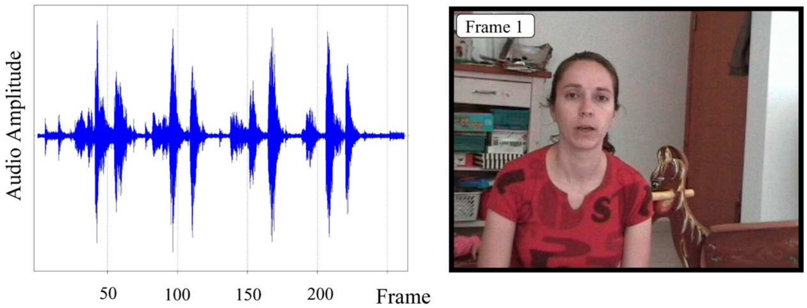


Fig. 6. Movie #2 includes a talking face and a moving wooden horse. [Left] The audio signal. [Right] A sample frame.

VIII. Experiments

The algorithm described is of a wide scope, handling localization of cross-modal correlations using insufficient data. In this section we demonstrate it for audio-visual analysis. In this domain, the signals are represented by some visual and audio features. Once the problem is solved, the results should be transferred from the feature space back into the image domain (pixels domain). The output of the localization algorithm is a weight $w_v(k)$ for each component k of the vector $\mathbf{v}$. The weights are transformed into an image $\mathbf{w}_v^{\text{Image}}$. For example, if wavelets are the domain of $\mathbf{v}$, then an inverse wavelet transform $\mathbf{w}_v^{\text{Image}} = \mathcal{W}^{-1}\mathbf{w}_v$ brings $\mathbf{w}_v$ to the pixel domain. Note that the image $\mathbf{w}_v^{\text{Image}}$ can have positive and negative components. We thus display the energy of the components, $e(\vec{x}) = |\mathbf{w}_v^{\text{Image}}(\vec{x})|^2$, where $\vec{x}$ is the pixel coordinate vector. Based on this energy distribution one can derive a measure of localization success. Details about such a criterion can be found in [26].

We now detail our experiments. Had we had only a single moving object in the field of view, its detection would have been trivial: standard image processing tools of motion detection would suffice. It would not require cross-modal analysis. Thus, to challenge the algorithm we deliberately based each of our experiments on two moving objects: only one of them is associated with the audio, while the other is a strong visual distraction (a rocking wooden-horse). Additional moving objects are expected to yield similar results. Moreover, in some experiments we added strong audio noises $(\text{SNR} = 1)$, in the form of unseen talking people, broadband noise, or background beats.

Our video sequences were sampled at 25 frames/s at resolution of 576 × 720 pixels.[12] The audio was sampled at 44.1 KHz. **Movie #1** features a hand playing a guitar and then a synthesizer. Such an example gives a good demonstration of *dynamics*. The hand playing motion is distracted by a rocking wooden-horse. Some raw data of this sequence appears in Fig. 1. **Movie #2** features a talking face and a distracting rocking wooden-horse as well. The audio plot and a representative frame of this sequence are shown in Fig. 6. Both movies can be linked through http://www.ee.technion.ac.il/~yoav/AudioVisual.html.

The experiments had the following features, aimed at demonstrating some capabilities of our approach:

- **Handling dynamics**. Each sequence was ≈10 s long. However, analysis was performed on intervals of $N_F = 32$ frames (≈1 second).

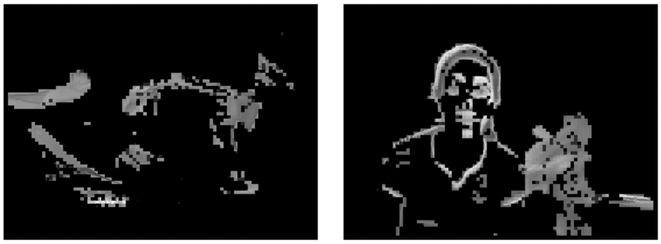


Fig. 7. Dynamic pixels expressed by the wavelet components in [left] Movie #1 and [right] Movie #2. Gray levels indicate the temporal average of pixel values. Black regions represent static pixels.

- **High spatial resolution (localization)**. In some of the prior work, pruning of visual features had been very aggressive, greatly decreasing spatio-temporal resolution. Our algorithm does *not* need this, thanks to the sparsity criterion. Nevertheless, memory limits currently restricted the number of visual features to $N_v = 3000$. The dynamic pixels in our frames were effectively represented by wavelet coefficients of such dimensions, as described below. The dynamic pixels are shown in Fig. 7. It is stressed that pruning was done only for reducing the computational load. However, we observed in experiments that using a larger number of features has a diminishing return. We aim to demonstrate high spatial resolution in the resulting visual localization.
- **No parameters to tweak**. The implementation has essentially no parameters. The selection of $N_F = 32$ represents our desire to localize brief events, but longer time intervals can be used as well. The selection of $N_v = 3000$ stems from hardware limits, but the results of our experiments observed robustness to this choice.
- **Simple audio representation**. Our experiments *did not attempt to filter sounds*, but rather to filter the visual signals. Hence, only a few audio bands were used. We analyzed the sequences using a single wide band $(N_a = 1)$, averaging sound energy at each frame (1/25th second). We then reanalyzed the data using $N_a = 4$ audio bands, selected as the strongest periodogram coefficients.

Since a sparse representation is desired, we worked on temporal-difference images, applying a wavelet transform to each of these difference-frames [45], [46]. We choose to use a wavelet decomposition of up to level 3. Coarser levels may incline the algorithm to choose coarser features. This reduces the ℓ^1 value but expands the spatial spread in the image domain.

[12] We used only the pixel intensities, and discarded the chromatic channels.

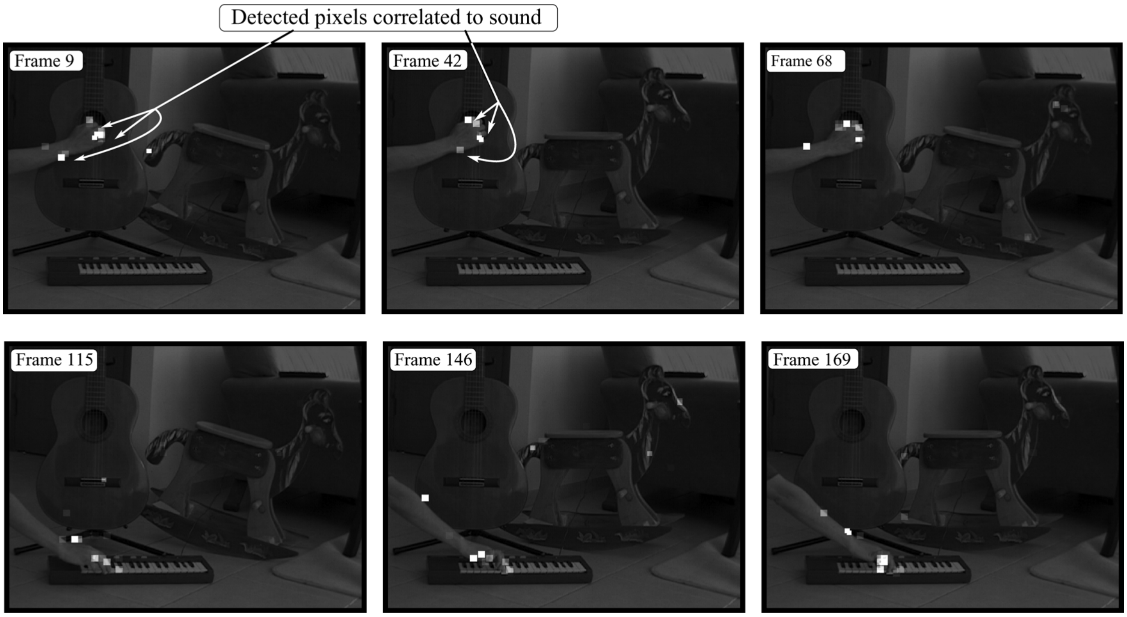


Fig. 8. The algorithm results, when run on Movie #1. For visualization, we overlayed the detected energy distribution with the corresponding sample raw frames. Localization concentrates on the playing fingers, which dynamically move from the guitar to the synthesizer. Sporadic detections exist in other areas, usually with much lower energies.

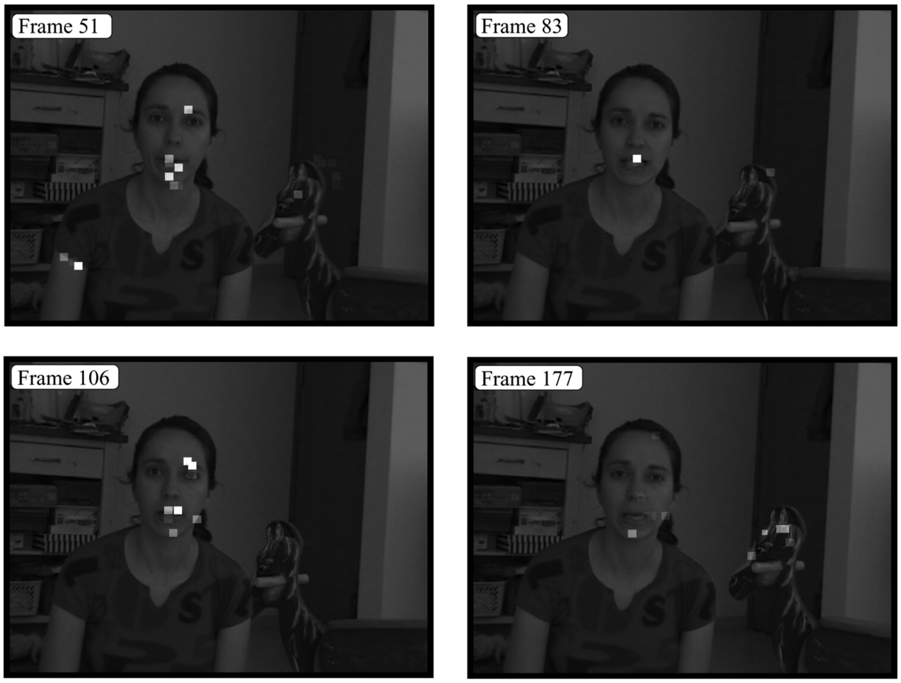


Fig. 9. Sample frames resulting from the algorithm, when run on Movie #2. The visualization is as described in Fig. 8. Localization in the mouth area is consistent. Sporadic detections exist in other areas, usually with much lower energies.

Fig. 8 shows sample frames resulting from the analysis of **Movie #1**. At each frame, we overlaid the energy distribution of the detected pixels $e(\vec{x})$ with the corresponding raw image. The algorithm pinpointed the source of the sound on the motion of the *fingers*, demonstrating both high spatial accuracy and temporal resolution. Compared to the large area occupied by dynamic pixels in Fig. 7, the detected pixels in Fig. 8 are concentrated in much smaller areas. Thus, high localization is achieved. Note that the algorithm handles *dynamics*. First, the guitar is detected, corresponding to its audio tones. When the hand played the synthesizer, the algorithm managed to shift its focus accordingly. The motion distractions (rocking horse) were successfully filtered out by our localization algorithm.

Similarly, Fig. 9 shows sample frames resulting from the analysis of **Movie #2**. Here pixels in the *mouth* were predominantly detected as correlated with the audio. Similarly to the

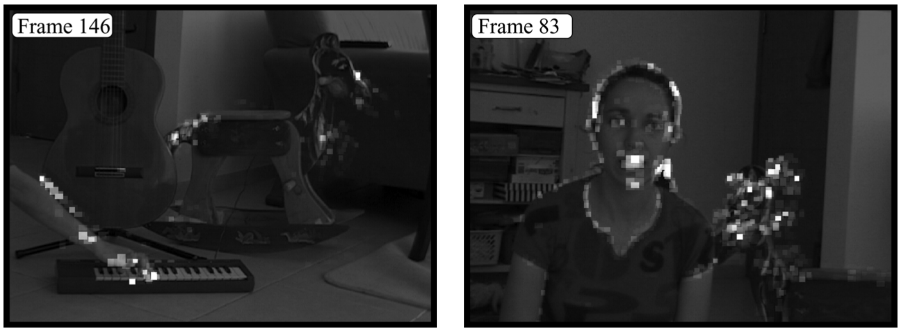


Fig. 10. Typical results of using ℓ^2 as a criterion. Compared to the corresponding frames shown in Figs. 8 and 9, the detected energy is much more spread, particularly in irrelevant areas (see the wrong detection of the horse on the right frame).

TABLE I
LOCALIZATION VALUES

	Using ℓ^1-norm	Using ℓ^2-norm
Movie #1	58 ± 20	4.0 ± 0.8
Movie #2	81 ± 20	2.9 ± 0.6

results of **Movie #1**, the motion distractions are successfully filtered out.

To judge the results, we compare our algorithm to the results of ℓ^2 regularization (19). Typical sample frames are shown in Fig. 10. They suffer from very poor localization and detection rate: there are many false-positives (especially detection of the moving horse), while the energy spreads over a large area. Note that in both experiments, the ℓ^2 and the ℓ^1 regularizations lead to full empirical correlation (4), $\hat{\rho} = 1$. Hence, the detected sets of features in Figs. 8 and 10 (say, in frame 146) are both considered as optimal CCA results. This is an example for the fact that CCA does not have a unique solution when data is insufficient. This example shows that the ℓ^1-norm leads to a sparse solution, which is consistent with our subjective expectation.

Table I reports the temporal mean and standard deviation of the empirical localization [26] values L_c, resulting from the use of either the ℓ^1 or ℓ^2-based localization algorithms. These quantitative results indicate that the ℓ^2-based solution achieves poor localization, compared to the ℓ^1-norm counterpart. As mentioned above, we repeated our experiments by sequentially adding three types of audio disturbances. The results were very similar to the ones reported in Table I, as well as visually. Moreover, we tested a multi-band audio representation using $N_a = 4$. The performance was very similar to that described in Figs. 8 and 9.

The experiments demonstrated elimination of irrelevant distractions (the rocking horse). If we had stationary sequences that were long enough, this elimination would have been achieved simply by CCA: the lack of correlation between the visual distraction and the audio would have been exposed. However, in the experiments, many irrelevant pixels may alias as correlated, since there is not enough data to reject them. The desired rejection of such irrelevant measurements was enabled by the sparsity-based regularization. In particular, the rejection of the rocking horse would occur also if it was spatially smaller, as long as it is not correlated to the guitar's audio. Had it been considered as part of the detected set in conjunction to the audio generating object, it would have increased $||\mathbf{w}_v||_0$ and $||\mathbf{w}_v||_1$ without increasing $\hat{\rho}$. On the other hand, if the guitar was strummed with the same rhythm as the rocking horse, then correlation might have existed between them. In such a case, the algorithm might detect pixels on the rocking horse, as well as on the hand area.

Interestingly, if N_F increases, the solution may be less sparse. This is typical to optimization problems: as data fitting becomes more reliable and prominent with added data, regularization effects become weaker.[13] Our priority is full correlation of the data, rather than sparsity. Hence, full empirical correlation is a constraint in our formulation. The sparsity prior only serves to regularize the solution.

IX. POSTPROCESSING FOR VISUALIZATION

The algorithm described above hardly exploits spatial coherence and temporal consistency, which are typical to audio-visual events. Still, it yields good results. Nevertheless, performance can be improved by further development of these aspects. This can be done by reformulating the optimization problem using priors expressing spatial coherence and temporal consistency. We opted for an alternative option, in which postprocessing is applied to the results of our algorithm, to filter out inconsistent behavior in time and space. This option is simpler and faster, since it involves concatenation of two relatively simple stages. As the post processing stage, we performed temporal median filtering (in windows of 10 frames), followed by spatial convolution with a 5×5 Gaussian kernel. The first step deletes temporal outliers, while the second stabilizes spatial positions and filters out fluctuations. Samples of resulting frames are shown in Fig. 11 and Fig. 12.

X. DISCUSSION

The algorithm presented here is parameter-free, and is thus robust to scenario variability. Nevertheless, the principles posed here can become the base for more elaborate localization approaches, that uses spatio-temporal consistency as a prior, as done in tracking methods. To enhance sparsity, it is possible to look for partial correlation, rather than full correlation. Such

[13] The other extreme is when data is almost nonexistent (say, only a single data point exists). Then, regularization dominates, leading to a meaningless solution to any optimization problem. In our problem, this would occur if $N_F = 1$. Then, the result would be a detection of an arbitrary single visual feature.

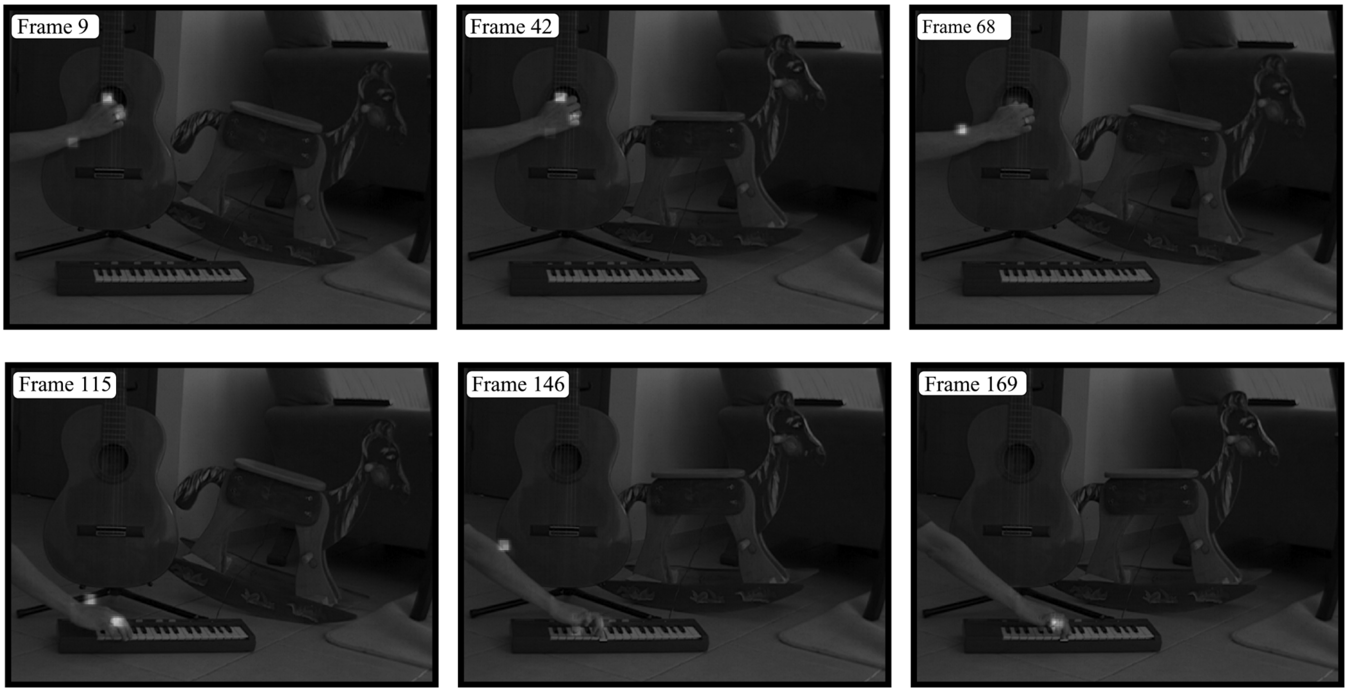


Fig. 11. Results of post processing of the algorithm output when the input is Movie #1. Compared to Fig. 8, the detected regions are much more stable and contain much less false-positives. Movie results are linked via http://www.ee.technion.ac.il/~yoav/AudioVisual.html.

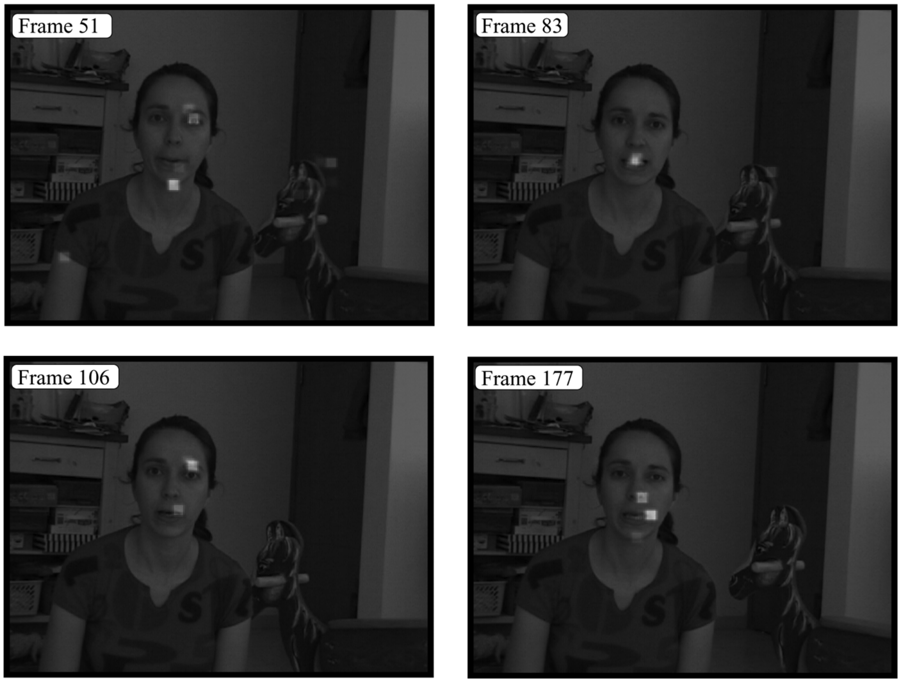


Fig. 12. Results of post processing of the algorithm output when the input is Movie #2. Compared to Fig. 9, the detected regions are much more stable and contain much less false-positives.

an algorithm may introduce parameters. Moreover, it is possible to solve the problem using other norms than those mentioned in this paper, to improve convergence. There are still open questions in this research, such as the nature of the method's breaking point. As in other computer vision and pattern recognition algorithms, introducing more complex scenarios would reveal new theoretical questions, which may lead to more complex methods.

It is possible to extend this approach, e.g., by a kernel version for treating nonlinear relations between the modalities [29], [31], [34]. In addition, time-lag between the modalities can be introduced as a variable in the optimization. In audio-visual analysis, this would enable estimation of object distances from the camera, based on the speed of sound. One may go further and generalize the problem formulation to multiple simultaneous events to be localized.

Our algorithm has dealt with correlation between two modalities, while one of the modalities has high dimensions (visual data) and the other has low dimensions (audio data). It is desirable to extend the approach, so that both modalities can be high dimensional. In addition, it would be interesting to consider extensions to more than two modalities. This may be based on generalized CCA [47]. This would be useful, for example, for correlating several synchronous cameras (multiple video channels) to audio. Furthermore, it is worth considering the application of our sparsity-based approach in other scientific domains that aim to correlate arrays of measurement vectors. These may include climatology, economy, sociology and medical research.

Appendix

A. Bounds of G

In this section of the appendix we prove that the penalty function given by (5) is bounded to the range [0,2]. Clearly (5) is nonnegative by definition. It may become zero when $\mathbf{V}\mathbf{w}_v = \mathbf{A}\mathbf{w}_a$, and hence the lower bound of G is zero as claimed. For the upper bound, we use the inequalities $\|\mathbf{y}-\mathbf{z}\|_2^2 \leq (\|\mathbf{y}\|_2 + \|\mathbf{z}\|_2)^2$ and $2\|\mathbf{y}\|_2\|\mathbf{z}\|_2 \leq \|\mathbf{y}\|_2^2 + \|\mathbf{z}\|_2^2$ (relation between geometric mean and algebraic mean). This yields

$$\begin{aligned}\frac{\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\mathbf{w}_a\|_2^2}{\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\mathbf{w}_a\|_2^2} &\leq \frac{\left(\|\mathbf{V}\mathbf{w}_v\|_2 + \|\mathbf{A}\mathbf{w}_a\|_2\right)^2}{\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\mathbf{w}_a\|_2^2}\\ &= 1 + \frac{2\|\mathbf{V}\mathbf{w}_v\|_2 \cdot \|\mathbf{A}\mathbf{w}_a\|_2}{\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\mathbf{w}_a\|_2^2}\\ &\leq 1 + \frac{\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\mathbf{w}_a\|_2^2}{\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\mathbf{w}_a\|_2^2}\\ &= 2. \qquad \text{(A-1)}\end{aligned}$$

This upper bound is tight, and is realized when $\mathbf{V}\mathbf{w}_v = -\mathbf{A}\mathbf{w}_a$. Thus, our claim that $0 \leq G \leq 2$ is proven. As explained in Section III, the range $0 \leq G \leq 1$ is equivalent to $1 \leq G \leq 2$.

B. Regularization by an Identity Matrix

In this section of the appendix we prove that regularization based on addition of the identity matrix, as posed in (24), leads to reduced correlation. Recall that for small enough ϵ, this regularization leads to the new penalty function

$$\widetilde{G}(\mathbf{w}_v) = G(\mathbf{w}_v) + \epsilon \frac{\|\mathbf{w}_v\|_2^2}{\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\|_2^2} \qquad \text{(B-1)}$$

where $G(\mathbf{w}_v)$ is the original penalty function of the non-regularized matrices. Let us define $\mathbf{w}_v^0$ as the minimizer of the regularized cost function $\widetilde{G}(\mathbf{w}_v)$. This minimizer must lead to a zero derivative, implying

$$\left[\frac{\partial}{\partial \mathbf{w}_v} \widetilde{G}(\mathbf{w}_v)\right]_{\mathbf{w}_v^0} = 0 \qquad \text{(B-2)}$$

leading to

$$\left[\frac{\partial}{\partial \mathbf{w}_v} G(\mathbf{w}_v)\right]_{\mathbf{w}_v^0} + 2\epsilon \mathbf{w}_v^0 \left(\|\mathbf{V}\mathbf{w}_v^0\|_2^2 + \|\mathbf{A}\|_2^2\right) - 2\epsilon \mathbf{V}^T\mathbf{V}\mathbf{w}_v^0 \left\|\mathbf{w}_v^0\right\|_2^2 = 0. \qquad \text{(B-3)}$$

Using the derivative of G given in (7) we obtain

$$\left[\epsilon \mathbf{w}_v^0 + \mathbf{V}^T\left(\mathbf{V}\mathbf{w}_v^0 - \mathbf{A}\right)\right] - \frac{\left\|\mathbf{V}\mathbf{w}_v^0 - \mathbf{A}\right\|_2^2 + \epsilon\left\|\mathbf{w}_v^0\right\|_2^2}{\left\|\mathbf{V}\mathbf{w}_v^0\right\|_2^2 + \|\mathbf{A}\|_2^2} \mathbf{V}^T\mathbf{V}\mathbf{w}_v^0 = 0. \qquad \text{(B-4)}$$

Let us assume first, that the solution $\mathbf{w}_v^0$ leads to complete coherence, and then show that this leads to a contradiction. Complete coherence (i.e., maximal correlation) implies $G(\mathbf{w}_v^0) = 0$, which is obtained if $\mathbf{V}\mathbf{w}_v^0 = \mathbf{A}$. Plugging this into (B-4), and using the fact that $\epsilon > 0$, yields

$$\frac{\left\|\mathbf{V}\mathbf{w}_v^0\right\|_2^2 + \|\mathbf{A}\|_2^2}{\|\mathbf{w}_v^0\|_2^2} \mathbf{w}_v^0 = \mathbf{V}^T\mathbf{V}\mathbf{w}_v^0. \qquad \text{(B-5)}$$

Multiplying (B-5) by $\mathbf{V}$, we get

$$\frac{\left\|\mathbf{V}\mathbf{w}_v^0\right\|_2^2 + \|\mathbf{A}\|_2^2}{\|\mathbf{w}_v^0\|_2^2} \mathbf{V}\mathbf{w}_v^0 = \mathbf{V}\mathbf{V}^T\mathbf{V}\mathbf{w}_v^0. \qquad \text{(B-6)}$$

Using $\mathbf{V}\mathbf{w}_v^0 = \mathbf{A}$ again, we obtain

$$\frac{\left\|\mathbf{V}\mathbf{w}_v^0\right\|_2^2 + \|\mathbf{A}\|_2^2}{\|\mathbf{w}_v^0\|_2^2} \mathbf{A} = \mathbf{V}\mathbf{V}^T\mathbf{A}. \qquad \text{(B-7)}$$

The last equation implies that the arbitrary audio data vector $\mathbf{A}$ is an eigenvector of the matrix $\mathbf{V}\mathbf{V}^T$. However, as $\mathbf{A}$ and $\mathbf{V}$ stem from distinct sources, they do not, in general, satisfy this property. Thus, the solution that minimizes $\widetilde{G}$ does not null G, i.e., the absolute correlation value is reduced.

C. Sparsity Using ℓ^1

Suppose we seek to solve (25). This task is highly complex (known to be NP-hard) [36], [37], being a combinatorial problem whose complexity grows exponentially with the number of columns in $\mathbf{V}$. Fortunately, we may use an approximation method that replaces the ℓ^0-norm with an ℓ^1 norm, yielding (26). This convex approximation is known as the *basis pursuit* algorithm [40]. The advantage of such a change is that it can be cast as a linear programming problem and be solved by modern interior point methods, even for very large N_v [40].

Recent studies have established that if the solution of (25) is sparse enough, then: i) no other solution exists with the same or lower cardinality (*uniqueness*); and ii) solving (26) yields a solution which is identical to the solution of (25) (*equivalence*) [36], [37]. Both the uniqueness and the equivalence results are derived from the properties of the matrix $\mathbf{V}$. Defining $\mathbf{v}_n$ as the nth column in this matrix, the *mutual coherence* is defined as

$$M = \max_{n \neq j} \frac{\left|\mathbf{v}_n^T \mathbf{v}_j\right|}{\|\mathbf{v}_n\|_2 \|\mathbf{v}_j\|_2} \qquad \text{(C-1)}$$

for $n, j = 1, 2, \ldots, N_v$. The work reported in [36]–[39] shows that uniqueness and equivalence of (25) and (26) hold true if the solution satisfies[14]

$$\left\|\mathbf{w}_v^{\text{optimal}}\right\|_0 < 0.5(1 + 1/M). \qquad \text{(C-2)}$$

[14] It can be shown [38] that $\sqrt{(N_v - N_F)/(N_F(N_v - 1))} \leq M \leq 1$.

In this case, the solution is considered to be a *highly sparse solution*, and solving (26) can replace (25).

The bound in (C-2) is rather restrictive. It is very conservative since it relates to worst-case scenarios. There are, however, cases where this restriction in meaningless. Consider an extreme case where the matrix $\mathbf{V}$ includes two identical columns. In this case, (C-1) yields $M = 1$, implying that uniqueness and equivalence hold true for $\mathbf{w}_v$ vectors having less than a single nonzero component (i.e., the entire vector is zero). Such an observation is useless. Apparently, empirical tests show that basis pursuit (26) recovers the solution of (25) for cases far exceeding the aforementioned bound.

Encouraged by these empirical observations, very recent theoretical analysis [42], [43], [48] addressed the above questions from a probabilistic point of view. This analysis has replaced a deterministic claim of "guaranteed uniqueness and equivalence" with a claim of "guaranteed uniqueness and equivalence with probability one." These studies establish a much higher bound on the cardinality of the solution to guarantee success.[15] These new results stand as supporting evidence to our experiments (Section VIII), where basis pursuit succeeded in locking on very sparse solutions.

[15] Those results assume a special structure of $\mathbf{V}$ and an asymptotic behavior, which do not necessary exist in our case.

References

[1] W. A. Landman and E. Klopper, "15-year simulation of the December to March rainfall season of the 1980s and 1990s using canonical correlation analysis (CCA)," *Water SA*, vol. 24, pp. 281–285, 1998.

[2] C. K. Wikle, "Spatio-temporal methods in climatology," in *Encyclopedia of Life Support Systems (EOLSS)*, A. H. El-Shaarawi and J. Jureckova, Eds. Oxford, U.K.: Eolss [Online]. Available: http://www.eolss.net, Developed Under the Auspices of the UNESCO.

[3] L. Navarro and E. M. Quilis, "Exploring the Spanish interbank yield curve," Instituto de Estudios Fiscales vol. P. T. N. 25/03, 2003.

[4] O. Friman, J. Cedefamn, P. Lundberg, M. Borga, and H. Knutsson, "Detection of neural activity in functional MRI using canonical correlation analysis," *Magn. Reson. Med.*, vol. 45, pp. 323–330, 2001.

[5] D. R. Hardoon, J. S. Taylor, and O. Friman, "KCCA for fMRI analysis," in *Proc. Med. Image Understanding Anal. (MIUA)*, 2004, pp. 141–144.

[6] C. Bregler and Y. Konig, "Eigenlips for robust speech recognition," in *Proc. IEEE ICASSP*, 1994, vol. 2, pp. 667–672.

[7] R. Cutler and L. Davis, "Look who's talking: speaker detection using video and audio correlation," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2000, vol. 3, pp. 1589–1592.

[8] M. Slaney and M. Covell, "Facesync: a linear operator for measuring synchronization of video facial images and audio tracks," *Adv. Neural Inf. Process. Syst.*, vol. 13, pp. 814–820, 2000.

[9] D. Murphy, T. H. Andersen, and K. Jensen, "Conducting audio files via computer vision," in *Proc. Gesture Workshop*, 2003, pp. 529–540.

[10] S. Deligne, G. Potamianos, and C. Neti, "Audio-visual speech enhancement with AVCDCN (audio-visual codebock dependent cepstral normalization)," in *IEEE Workshop on Sensor Array and Multichannel Signal Process.*, 2002, pp. 68–71.

[11] J. W. Fisher, III, T. Darrell, W. Freeman, and P. Viola, "Learning joint statistical models for audio-visual fusion and segregation," *Adv. Neural Inf. Process. Syst.*, vol. 13, pp. 772–778, 2001.

[12] J. Hershey and J. Movellan, "Audio-vision: Using audio-visual synchrony to locate sound," *Adv. Neural Inf. Process. Syst.*, vol. 12, pp. 813–819, 1999.

[13] D. Li, N. Dimitrova, M. Li, and I. K. Sethi, "Multimedia content processing through cross-modal association," in *Proc. ACM Int. Conf. Multimedia*, 2003, pp. 604–611.

[14] H. J. Nock, G. Iyengar, and C. Neti, "Assessing face and speech consistency for monologue detection in video," in *Proc. ACM Int. Conf. Multimedia*, 2002, pp. 303–306.

[15] P. Smaragdis and M. Casey, "Audio/visual independent components," in *Int. Symp. ICA BSS*, 2003, pp. 709–714.

[16] M. Song, J. Bu, C. Chen, and N. Li, "Audio-visual based emotion recognition—a new approach," *Proc. IEEE Comp. Vis. Pattern Recognit.*, vol. 2, pp. 1020–1025, 2004.

[17] J. Driver, "Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading," *Nature*, vol. 381, pp. 66–68, 1996.

[18] D. E. Feldman and E. I. Knudsen, "An anatomical basis for visual calibration of the auditory space map in the barn owl's midbrain," *J. Neurosci.*, vol. 17, pp. 6820–6837, 1996.

[19] Y. Gutfreund, W. Zheng, and E. I. Knudsen, "Gated visual input to the central auditory system," *Science*, vol. 297, pp. 1556–1559, 2002.

[20] M. J. Beal, N. Jojic, and H. Attias, "A graphical model for audiovisual object tracking," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, pp. 828–836, 2003.

[21] B. Kapralos, M. R. M. Jenkin, and E. Milios, "Audiovisual localization of multiple speakers in a video teleconferencing setting," *Int. J. Imag. Syst. Technol.*, vol. 13, pp. 95–105, 2003.

[22] C. Schauer and H. M. Gross, "A computational model of early auditory-visual integration," in *Proc. Pattern Recognit. Symp. Lecture Notes in Comput. Sci.*, 2003, vol. 2781, pp. 362–369.

[23] J. Vermaak, M. Gangnet, A. Blake, and P. Perez, "Sequential Monte Carlo fusion of sound and vision for speaker tracking," in *Proc. IEEE Int. Conf. Comp. Vis.*, 2001, vol. 1, pp. 741–746.

[24] J. W. Fisher, III and T. Darrell, "Speaker association with signal-level audiovisual fusion," *IEEE Trans. Multimedia*, vol. 6, pp. 406–413, 2004.

[25] S. Shwartz, M. Zibulevsky, and Y. Y. Schechner, "Fast kernel entropy estimation and optimization," *Signal Process.*, vol. 85, pp. 1045–1058, 2005.

[26] E. Kidron, Y. Y. Schechner, and M. Elad, "Pixels that sound," *Proc. IEEE Comp. Vis. Pattern Recognit.*, vol. 1, pp. 88–96, 2005.

[27] H. Knutsson, M. Borga, and T. Landelius, "Learning Canonical Correlations," Comput. Vis. Lab., Linköping Univ., Sweden, Tech. Rep. LiTH-ISY-R-1761, S-581 83, 1995.

[28] ——, "Learning multidimensional signal processing," in *Proc. Int. Conf. Pattern Recognit.*, 1998, vol. II, pp. 1416–1420.

[29] F. Bach and M. Jordan, "Kernel independent component analysis," *J. Mach. Learn. Res.*, vol. 3, pp. 1–48, 2002.

[30] T. D. Bie and B. D. Moor, "On the regularization of canonical correlation analysis," in *Int. Symp. ICA BSS*, 2003, pp. 785–790.

[31] T. Melzer, M. Reiter, and H. Bischof, "Appearance models based on kernel canonical correlation analysis," *Pattern Recognit.*, vol. 36, pp. 1961–1971, 2003.

[32] K. D. Cock and B. D. Moor, "Subspace angles and distances between ARMA models," *Syst. Contr. Lett.*, vol. 46, pp. 265–270, 2002.

[33] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 3rd ed. Baltimore, MD: The Johns Hopkins Univ. Press, 1996.

[34] L. Wolf and A. Shashua, "Learning over sets using kernel principal angles," *J. Mach. Learn. Res.*, vol. 4, pp. 913–931, 2003.

[35] G. Farnebäck, "A unified framework for bases, frames, subspace bases, and subspace frames," in *Proc. Scand. Conf. Image Anal.*, 1999, pp. 341–349.

[36] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization," *Proc. Nat. Acad. Sci.*, vol. 100, pp. 2197–2202, 2003.

[37] R. Gribonval and M. Nielsen, "Sparse representations in unions of bases," *IEEE Trans. Inf. Theory*, vol. 49, pp. 3320–3325, 2003.

[38] D. L. Donoho and M. Elad, "On the stability of the basis pursuit in the presence of noise," *EURASIP Signal Process. J.*, vol. 86, pp. 511–532, 2006.

[39] D. L. Donoho, M. Elad, and V. Temlyakov, "Stable recovery of sparse overcomplete representations in the presence of noise," *IEEE Trans. Inf. Theory*, vol. 52, pp. 6–18, 2006.

[40] S. Chen, D. Donoho, and M. Saunders, "Atomic decomposition by basis pursuit," *SIAM Rev.*, vol. 43, pp. 129–159, 2001.

[41] M. Elad and A. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," *IEEE Trans. Inf. Theory*, vol. 48, pp. 2558–2567, 2002.

[42] E. Candes, J. Romberg, and T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, pp. 489–509, 2006.

[43] D. L. Donoho, "For most large underdetermined systems of linear equations, the minimal ℓ^1-norm solution is also the sparsest solution," Statistics Dept., Stanford Univ., Stanford, CA, Tech. Rep., 2004.

[44] I. F. Gorodnitsky and B. D. Rao, "Sparse signal reconstruction from limited data using FOCUSS: A re-weighted norm minimization algorithm," *IEEE Trans. Signal Process.*, vol. 45, pp. 600–616, 1997.

[45] D. L. Donoho and A. G. Flesia, "Can recent innovations in harmonic analysis explain key findings in natural image statistics?," *Network: Comput. Neural. Syst.*, vol. 12, pp. 371–393, 2001.

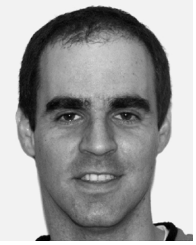
[46] S. G. Mallat, "A theory multiresolution signal decomposition: the wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, pp. 674–693, 1989.

[47] J. Kettenring, "Canonical analysis of several sets of variables," *Biometrika*, vol. 58, pp. 433–451, 1971.

[48] M. Elad, "Sparse representations are most likely to be the sparsest possible," *J. Appl. Signal Process.*, vol. 2006, pp. 1–12, 2006.

Einat Kidron received the B.Sc and M.Sc. degrees from the Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, in 1996 and 2006, respectively.

She is a research and development engineer. From 1998 to 2004, she has been working on satellite communication and later on UWB wireless. Since 2004, she has been working in the field of signal and image processing, focusing on image enhancement and image restoration for mobile phone cameras.



Yoav Y. Schechner received the B.A. and M.Sc. degrees in physics and the Ph.D. degree in electrical engineering from the Technion-Israel Institute of Technology, Haifa, in 1990, 1996, and 1999, respectively.

During 2000–2002, he was a research scientist with the Computer Science Department, Columbia University, New York. Since 2002, he has been a faculty member with the Department of Electrical Engineering, Technion, where he heads the Hybrid Imaging Lab. His research is focused on computer vision and multimodal sensing.

Dr. Schechner was the recipient of the Wolf Foundation Award for Graduate Students in 1994, the Guttwirth Special Distinction Fellowship in 1995, the Israeli Ministry of Science (Eshkol) Distinction Fellowship and the Ollendorff Award in 1998, the Scwartz Foundation Award in 1999, and the Morin Fellowship in 2000–2002. He is now a Landau Fellow—supported by the Taub Foundation, and an Alon Fellow.

Michael Elad received the B.Sc., M.Sc., and D.Sc. degrees from the Department of Electrical Engineering, Technion-Israel Institute of Technology, Haifa, in 1986, 1988, and 1997, respectively.

From 1988 to 1993, he served in the Israeli Air Force. From 1997 to 2000, he worked at Hewlett-Packard Laboratories as an R&D engineer. From 2000 to 2001, he headed the Research Division at Jigami Corporation, Israel. During 2001–2003, he was a Research Associate with the Computer Science Department, SCCM program, Stanford University, Stanford, CA. Since September 2003, he has been with the Department of Computer Science, Technion, as an Assistant Professor. His interests include the field of signal and image processing, specializing in particular on inverse problems, sparse representations, and overcomplete transforms.

Dr. Elad received the Technion's Best Lecturer Award four times (1999, 2000, 2004, and 2005). He is also the recipient of the Guttwirth and the Wolf fellowships. He is currently serving as an Associate Editor for the IEEE Transactions on Image Processing and EURASIP Signal Processing journals.