Pixels that Sound

Einat Kidron and Yoav Y. Schechner
Dept. Electrical Engineering
Technion - Israel Inst. Technology
Haifa 32000, ISRAEL
einatt@tx.technion.ac.il
yoav@ee.technion.ac.il

Michael Elad
Dept. Computer Science
Technion - Israel Inst. Technology
Haifa 32000, ISRAEL
elad@cs.technion.ac.il

Abstract

People and animals fuse auditory and visual information to obtain robust perception. A particular benefit of such cross-modal analysis is the ability to localize visual events associated with sound sources. We aim to achieve this using computer-vision aided by a single microphone. Past efforts encountered problems stemming from the huge gap between the dimensions involved and the available data. This has led to solutions suffering from low spatio-temporal resolutions. We present a rigorous analysis of the fundamental problems associated with this task. Then, we present a stable and robust algorithm which overcomes past deficiencies. It grasps dynamic audio-visual events with high spatial resolution, and derives a unique solution. The algorithm effectively detects pixels that are associated with the sound, while filtering out other dynamic pixels. It is based on canonical correlation analysis (CCA), where we remove inherent ill-posedness by exploiting the typical spatial sparsity of audio-visual events. The algorithm is simple and efficient thanks to its reliance on linear programming and is free of user-defined parameters. *To quantitatively assess the performance, we devise a localization criterion. The algorithm capabilities were demonstrated in experiments, where it overcame substantial visual distractions and audio noise.*

1 Introduction

There is a growing interest in multi-sensor processing. A particularly interesting sensor combination involves visual motion in conjunction with associated *audio*. Activity in computer vision involving audio analysis has various research aspects [4, 26], including lip reading [3, 25], analysis and synthesis of music from motion [22], audio filtering based on motion [6], and source separation based on vision [14, 17, 20, 23, 27]. We note that physiological evidence and analysis of biological systems show that fusion of audio-visual information is used to enhance perception [9, 12, 16].

In this work, we focus on accurately *pinpointing the visual localization* of 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 [3, 4, 20, 23] or other specific classes of sources [22], but seek a general and effective algorithm to achieve this goal. Some existing methods use several microphones (emulating binaural hearing), where stereo triangulation indicates the spatial location of the sources [2, 24, 28]. In contrast, we seek a very sharp spatial localization of the sound source, using a single microphone (monaural hearing) and a video stream. Moreover, we wish the localization method to perform well, even if interfering sounds exist, unrelated to the desired object.

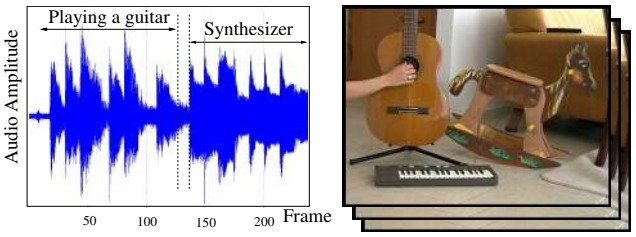

Figure 1. Audio data [Left] is sequential, requiring $\mathcal{O}(10^4)$ samples/sec. Corresponding video [Right] frames are highly parallel (multi-pixel), requiring $\mathcal{O}(10^7)$ samples/sec. Pinpointing a sound source in the images by correlation requires dimensionality reduction of the visual signal. This reduction involves too many degrees of freedom.

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 (1D) subspace [20, 25, 27]. Thus, two 1D variables represent the audio and the visual signals. Localization algorithms typically seek 1D representations that best correlate [17, 20, 25]. However, as shown in this paper, 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 [3, 4, 25] to reduce the number of unknowns. Others consider acquisition of very long sequences to ensure sufficient data quantities [6, 20]. 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 modal representations [13], while trading off ℓ^2-based regularization terms. This approach requires multiple tune-up parameters, and suffers from the complexity of estimating MI using Parzen windows. While MI better indicates cross-modal statistical dependency, there is no guarantee for a unique solution, due to the non-convexity of MI.

In this paper we describe an algorithm that overcomes all those difficulties. It results in high spatio-temporal localization, which is unique and stable. 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-though the analysis is based on very short time intervals. The resulting sparsity does not compromise at all 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. To analyze performance, we propose a quantitative criterion for the visual localization of sounds. We then demonstrate the merits of the algorithm in experiments using real data.

2 Canonical Correlation: Limitations

An important tool for understanding the relationship between sound and video is *canonical correlation analysis* (CCA). In this section we describe CCA, and the reason for its importance. We then indicate a fundamental limitation of that method in the context of our problem.

CCA deals with the correlation between two sets of random variables. The sets can be of different nature, 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 wavelet coefficients. Let $\mathbf{a}$ represent a corresponding audio signal, e.g., by the intensity of different sound bands (a temporal slice of the periodogram). 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 random variables, $\mathbf{v}^T\mathbf{w}_v$ and $\mathbf{a}^T\mathbf{w}_a$, where T denotes transposition. The correlation coefficient of these two variables defines the canonical correlation [19] between $\mathbf{v}$ and $\mathbf{a}$,

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

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.

Maximizing the data correlation seeks the subspaces $\mathbf{w}_v$ and $\mathbf{w}_a$ that optimize Eq. (1). Fortunately, this optimization problem can usually be solved easily. The reason is that it is equivalent to the following eigenvalue problem [19]:

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

Maximizing the absolute correlation is equivalent to finding the largest eigenvalue and its corresponding eigenvector. Inspecting the optimal $\mathbf{w}_v$, the components which have the largest magnitude indicate the visual components that best correlate with the projection of $\mathbf{a}$, and vice-versa.

At first sight, CCA may appear as a good tool for correlating audio to 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 of projections of feature vectors have been the core of several audio-visual algorithms [14, 17, 20, 25]. However, CCA and its related methods [20] have a serious shortcoming. The fundamental problem is the *scarcity of data* available in short time intervals, which is *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. For example, $\mathbf{C}_{vv}$ is estimated as the empirical matrix

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

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 make $\widehat{\mathbf{C}}_{vv}$ in Eq. (3) full rank, 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 this creates data shortage. The matrix $\widehat{\mathbf{C}}_{vv}$ becomes hugely rank deficient, hence (2) cannot be solved, making CCA ill-posed. This rank deficiency can be technically bypassed by regularization, e.g., by weighted averaging of $\widehat{\mathbf{C}}_{vv}$ with an identity matrix [1, 5, 21]. Such operations do not overcome the fundamental problem of unreliable statistics. They yield an arbitrary solution, which somewhat compromises the correlation ρ.

The gap between the amount of data and degrees of freedom is not limited to CCA. It affects methods based on MI as well [13]. Hence, very small images $\mathcal{O}(50\times 50)$ have been commonly used [3, 20, 23, 25], out of which only a few dozen features were selected by aggressive pruning or face detection steps (the latter limiting audio analysis to speech). In contrast, we seek localization of general unknown audio-visual sources, while handling intricate details and motion. In the following, we show how this can be achieved.

[1] Each of the vectors $\mathbf{v}$ and $\mathbf{a}$ is assumed to have a zero mean.

3 Sparsity: A Key to Alleviate Ill-Posedness

In this section we derive the set of solutions to audio-visual correlation, for cases where the temporal resolution is too short to acquire sufficient data. It is shown that this set is infinite. We then describe our approach, which leads to a unique solution based on a spatial sparsity criterion.

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. Defining $\widehat{\mathbf{C}}_{vv} = \mathbf{V}^T\mathbf{V}$, $\widehat{\mathbf{C}}_{aa} = \mathbf{A}^T\mathbf{A}$ and $\widehat{\mathbf{C}}_{va} = \mathbf{V}^T\mathbf{A}$, the empirical canonical correlation[2] (Eq. 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 $|\hat{\rho}|$. Note that there exists an alternative formulation to CCA, called principal angles [29], which is the constraint optimization

$$\begin{array}{ll} \max_{\mathbf{w}_a,\mathbf{w}_v} & \{\mathbf{w}_v^T\mathbf{V}^T\mathbf{A}\mathbf{w}_a\} \\ \text{subject to} & \|\mathbf{V}\mathbf{w}_v\|^2 = 1,\ \|\mathbf{A}\mathbf{w}_a\|^2 = 1 \ . \end{array} \tag{5}$$

In [18] we prove that maximizing $|\hat{\rho}|$ is equivalent to unconstrained minimization of the objective function

$$G(\mathbf{w}_v, \mathbf{w}_a) = \frac{\|\mathbf{V}\mathbf{w}_v - \mathbf{A}\mathbf{w}_a\|_2^2}{\|\mathbf{V}\mathbf{w}_v\|_2^2 + \|\mathbf{A}\mathbf{w}_a\|_2^2} \tag{6}$$

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

$$(1-G)\widehat{\mathbf{C}}_{vv}\mathbf{w}_v = \widehat{\mathbf{C}}_{va}\mathbf{w}_a \ , \ (1-G)\widehat{\mathbf{C}}_{aa}\mathbf{w}_a = \widehat{\mathbf{C}}_{av}\mathbf{w}_v \tag{7}$$

The equations in (7) yield CCA equations [19]. Particularly, if $\widehat{\mathbf{C}}_{vv}$ and $\widehat{\mathbf{C}}_{aa}$ are invertible, Eq. (7) leads to Eq. (2), for $\hat{\rho}^2 = (1-G)^2$. Hence, $G = 0$ is equivalent[3] to $|\hat{\rho}| = 1$. For this reason, minimizing G, when $G \leq 1$, maximizes $|\hat{\rho}|$.

We now progress by first looking at cases where $N_a = 1$, i.e., the audio is characterized by a single feature. In Sec. 3.2 we extend the analysis to multiple audio bands.

3.1 A Single Audio Band

When $N_a = 1$ we may set the scalar $\mathbf{w}_a$ to the value 1, since the penalty function in (6) is scale invariant (multiplying $\mathbf{w}_v$ and $\mathbf{w}_a$ by the same constant does not change the function's value). We still need to find the optimal $\mathbf{w}_v$ for minimizing G. For the moment, let us concentrate on minimizing the numerator of Eq. (6)

$$g(\mathbf{w}_v) = \|\mathbf{V}\mathbf{w}_v - \mathbf{A}\|_2^2 \ . \tag{8}$$

As shown later, the denominator is usually unimportant.

Suppose for a moment that a vector $\mathbf{w}_v$ exists such that $g(\mathbf{w}_v) = 0$. This vector minimizes $G(\mathbf{w}_v, \mathbf{w}_a)$ since the denominator of Eq. (6) is necessarily non-zero.[4] Moreover, this solution yields complete coherence, $|\hat{\rho}| = 1$, as desired. However, we now show that this estimation is ill-posed. Requiring $g(\mathbf{w}_v) = 0$ implies

$$\mathbf{V}\mathbf{w}_v = \mathbf{A} \ . \tag{9}$$

Since $N_a = 1$, $\mathbf{A}$ is a *column* vector of length N_F. As discussed in Sec. 2, $N_v \gg N_F$, where N_v is the length of $\mathbf{w}_v$. Therefore, in the set of linear equations (9), the number of equations is much smaller than the number of unknowns. Hence, the number of possible solutions is *infinite*. Due to the scarce data, there are infinite distributions of visual features which 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 $\mathrm{rank}(\mathbf{V}) = N_F$, guaranteeing 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.[5] In fact, noise in the visual data guarantees this outcome. However, visual noise implies strong correlation of "junk" features to the audio.

Underdetermined problems are commonly regularized by prefering the minimal energy solution [11]. In our case this would be

$$\min \|\mathbf{w}_v\|_2 \quad \text{subject to} \quad \mathbf{V}\mathbf{w}_v = \mathbf{A} \ . \tag{10}$$

In the context of the audio-visual problem, this results in visual *poor 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 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 this criterion is demonstrated in the experiments shown in Sec. 6.

To overcome this problem, we translate the locality assumption to a requirement that the sought solution should be *sparse*.[6] Our goal is that the optimal solution will have a minimal number of components. Thus, out of the entire space of possible correlated projections, we aim to solve:

$$\min \|\mathbf{w}_v\|_0 \quad \text{subject to} \quad \mathbf{V}\mathbf{w}_v = \mathbf{A} \ , \tag{11}$$

[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 Eq.(4), and is thus discarded throughout the paper.

[3] Also $G = 2$ leads to $|\hat{\rho}| = 1$. In [18] we show that 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, we can focus on minimizing G towards zero.

[4] This is true since $\mathbf{w}_a = 1$ and $\mathbf{A}$ is a non-zero vector.

[5] The case where $g(\mathbf{w}_v)$ has no zero is treated in [18], with similar conclusions to the ones described here.

[6] Sparsity is enhanced using a wavelet representation of temporal-difference images.

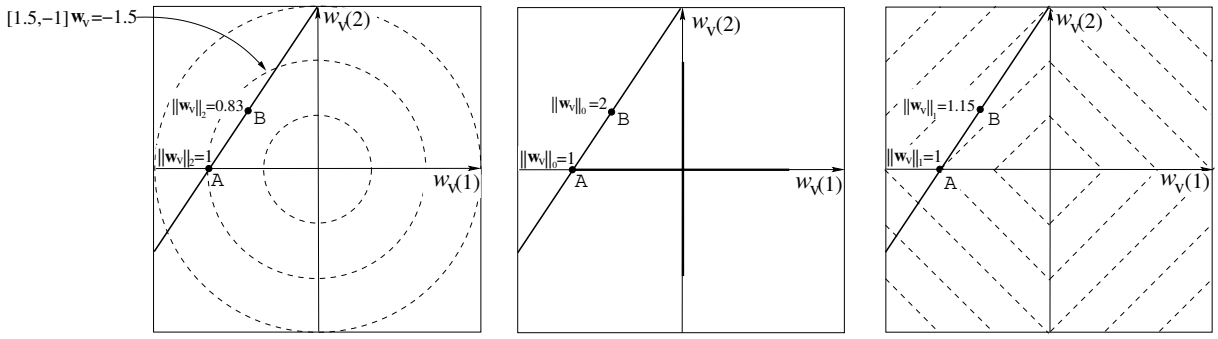

Figure 2. A 2D example of optimization under [Left] ℓ^2-norm [Middle] ℓ^0-norm [Right] ℓ^1-norm. The dashed contours represent iso-norm levels. On the linear constraint $\mathbf{V}\mathbf{w}_v = \mathbf{A}$ (solid line), point B minimizes $\|\mathbf{w}_v\|_2$, but it has substantial energy in all components. 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.

where $\|\cdot\|_0$ is the ℓ^0-norm of a vector space (the number of non-zero vector coefficients). In the simple example depicted in the middle of Fig. 2, the optimal solution according to this criterion (point A) has a single component out of two. Unfortunately, this criterion is not convex, and the complexity of its optimization is exponential [7, 15] in N_v.

We bypass this difficulty by convexizing the problem and solving

$$\min \|\mathbf{w}_v\|_1 \quad \text{subject to} \quad \mathbf{V}\mathbf{w}_v = \mathbf{A} \ , \tag{12}$$

where ℓ^1 is used instead of ℓ^0. In the right part of Fig. 2, the solution optimizing this criterion has a single component (point A), just as under the ℓ^0 criterion. All other points in the linear constraint $\mathbf{V}\mathbf{w}_v = \mathbf{A}$ have a larger ℓ^1-norm. Moreover, this figure shows the convexity of the ℓ^1 criterion.

In general, the equivalence of the ℓ^0 and ℓ^1 problems (11,12) has been studied in depth during the last couple of years from a pure mathematical perspective. Preliminary contributions in this direction considered deterministic sufficient conditions for this equivalence [7, 15]. More recently, a probabilistic approach has been introduced, showing that equivalence holds true far beyond the limits determined by these sufficient conditions [10]. Owing to this theoretical progress, formulating sparsity using the ℓ^1-norm is reliable.

The formulation in Eq. (12) can be posed as a *linear programming* problem, and thus can be solved *efficiently* even for a very large N_v. Moreover, the solution is *unique* because of the convexity of the formulation (except for special cases discussed in Sec. 4). Eq. (12) influences the solution energy to concentrate on few visual features which strongly correlate with the audio. It penalizes for dispersed components, such as the random "junk" features described above (e.g., image noise).

3.2 Multiple Audio Bands

We now generalize the single-band analysis of Sec. 3.1 to audio signals that are divided into multiple bands. We analyze here the scenario in which the cost function G has a zero value. This allows us to concentrate on the numerator of Eq. (6). The numerator is zero if and only if

$$\mathbf{V}\mathbf{w}_v = \mathbf{A}\mathbf{w}_a \ . \tag{13}$$

As before, if $\mathrm{rank}(\mathbf{V}) = N_F$, a zero solution of G is guaranteed. As claimed in Sec. 3.1, this is a highly probable event, especially for noisy visual data. 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. This case is treated in [18].

Similarly to Sec. 3.1, Eq. (13) 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 audio ℓ^1-ball, $\|\mathbf{w}_a\|_1 = 1$. However, the set $\|\mathbf{w}_a\| = 1$ is not convex. To keep enjoying the benefits of a convex problem formulation, the following process is performed. 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

$$\begin{aligned} s_q = \min \|\mathbf{w}_v\|_1 \quad & \text{subject to} \\ & \left\{\mathbf{V}\mathbf{w}_v = \mathbf{A}\mathbf{w}_a \ , \ \mathbf{h}_q^T\mathbf{w}_a = 1 \ , \ \mathbf{H}_q\mathbf{w}_a \geq 0\right\} \end{aligned} \tag{14}$$

where $\mathbf{h}_q$ is a vector 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, Eq. (14) 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 is done by running Eq. (14) over all[7] values of q, and then selecting the optimal q by

$$\hat{q} = \arg\min s_q \ . \tag{15}$$

[7] 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.

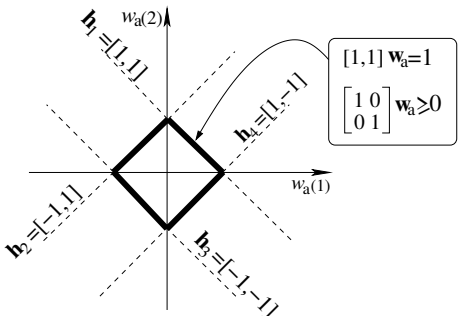

Figure 3. A 2D illustration of the faces of the ℓ^1-ball in the audio space.

The unique vectors $\mathbf{w}_v$ and $\mathbf{w}_a$ which we seek are then derived by using this specific $\hat{q}$ in Eq. (14). 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.

4 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? We claim that 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. Thus, we term this phenomenon as the *chorus ambiguity*.

Our algorithm (12,14,15) 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, this implies that for this special scenario, Eq. (12) does not have a unique solution, but rather a set of them.[8] Still, this effect does not hinder the optimization process: the linear programming converges to one of those solutions, depending on the initialization.

5 Quantitative Localization Criterion

Sec. 3 describes how to uniquely solve the audio-visual correlation problem. We now describe how the results are translated to the image (pixels) domain, and how their performance can be judged. 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 $\tilde{\mathbf{w}}_{\tilde{v}}$. For example, if wavelets are the domain of $\mathbf{v}$, then an inverse wavelet transform of $\mathbf{w}_v$ brings it to the pixel domain:

$$\tilde{\mathbf{w}}_{\tilde{v}} = \mathcal{W}^{-1}\mathbf{w}_v \quad . \tag{16}$$

Note that the image $\tilde{\mathbf{w}}_{\tilde{v}}$ can have positive and negative components. We thus display the energy of the components:

$$e(\vec{x}) = |\tilde{w}_{\tilde{v}}(\vec{x})|^2 \quad , \tag{17}$$

where $\vec{x}$ is the pixel coordinate vector. This energy distribution forms the basis for a localization criterion. High localization is obtained if most of the energy of the image $e(\vec{x})$ is concentrated in small areas that are *correct*.

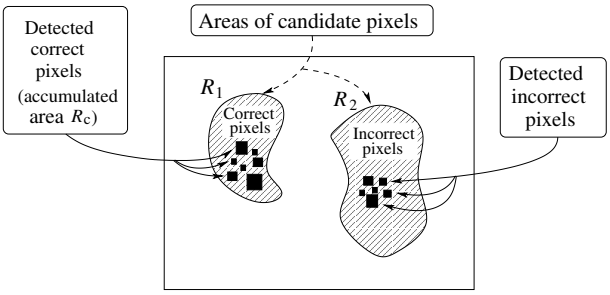

Figure 4. The candidate dynamic pixels occupy areas R_1 and R_2. Some of them are detected by the audio-visual localization algorithm (marked here in black). If detection is based on a multiresolution representation, then the area of detected pixels typically comprises of blocks of several fixed sizes.

Before audio-visual localization is attempted, all the dynamic pixels are *candidates* for detection. In Fig. 4, they are depicted as residing in regions R_1 and R_2. It must be stressed that all the pixels in those regions are dynamic, since pixels having values with negligible temporal variation are excluded. The pixels detected by the localization algorithm have $e(\vec{x}) > 0$. Some of them are in irrelevant areas. We determine a *correct* detection by manually defining R_1 as the area (of dynamic pixels) corresponding to the sound. For instance, in the sequence appearing in Fig. 1, R_1 includes only pixels in which the hand is moving. The set of correctly detected pixels

$$\mathcal{D}_c \doteq \{\vec{x} \,:\, e(\vec{x}) > 0 \text{ and } \vec{x} \in R_1\} \tag{18}$$

occupies a cumulative area R_c. The localization criterion is

$$L_c = \frac{\sum_{\vec{x}\in\mathcal{D}_c} e(\vec{x})}{\sum_{\vec{x}} e(\vec{x})} \cdot \frac{R_1 + R_2}{R_c} \quad . \tag{19}$$

It can be easily seen that if there is no preference for localization at the correct region, then $L_c = 1$. The case where $L_c < 1$ indicates failure, as most of the energy is outside the correct region. We seek $L_c \gg 1$, meaning that the energy is concentrated in small areas of correct identity.

[8] The $\|\mathbf{w}_v\|_0$ criterion locks exclusively into any single person in the chorus, while the $\|\mathbf{w}_v\|_1$ 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 [7, 15].

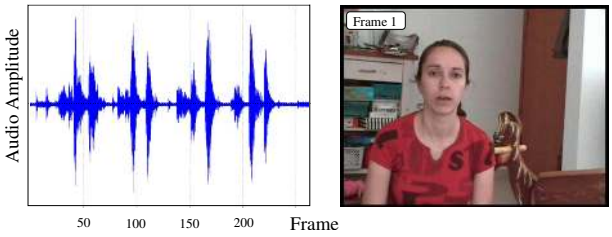

Figure 5. Movie #2 includes a talking face and a moving wooden horse. [Left] The audio signal. [Right] A sample frame.

Figure 6. Dynamic pixels expressed by the wavelet coefficients in [Left] Movie #1 [Right] Movie #2. Graylevels indicate the temporal average of pixels values. Black regions represent static pixels.

6 Experiments

In this section we present results of experiments based on real video sequences. The sequences were sampled at 25 frames/sec at resolution of 576×720 pixels.[9] The audio was sampled at 44.1KHz. **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. 5. Both movies can be linked through *http://www.ee.technion.ac.il/∼ yoav/AudioVisual.html* .

The experiments had the following features, aimed at demonstrating the strength of our algorithm:

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

• **Handling false-positives and noise.** The sequences deliberately include strong visual distractions (a rocking wooden-horse), challenging the algorithm. Moreover, in some experiments we sequentially added strong audio noises (SNR=1), in the form of unseen talking people (via a recording), broadband noise, or background beats.

• **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 components of such dimensions, as described below. The dynamic pixels are shown in Fig. 6. It is stressed that pruning was done only for reducing the computational load. Yet, we aim to demonstrate high spatial resolution in the resulting visual localization.

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

• **No parameters to tweak.** The implementation has essentially no parameters (e.g. weights of priors). 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 are robust to this choice, as verified in experiments.

• **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 ($1/25$'th second). We then re-analyzed 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. A wavelet transform was applied to each of these difference-frames [8]. For very sharp localization, we need to retain the image details. We thus performed wavelet decomposition only into levels corresponding to high resolution (up to level 3). Coarser levels can be used, but may hinder high localization.

Fig. 7 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. 6, the detected pixels in Fig. 7 are concentrated in much smaller areas. Thus, high localization was achieved. Note that the algorithm handled the *dynamics*. First, the guitar was detected, corresponding its audio tones. When the hand played the synthesizer, the algorithm managed to shift its focus. The motion distractions (rocking horse) were successfully filtered out by our audio-visual localization algorithm.

Similarly, Fig. 8 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 results of Movie #1, the motion distractions are successfully filtered out.

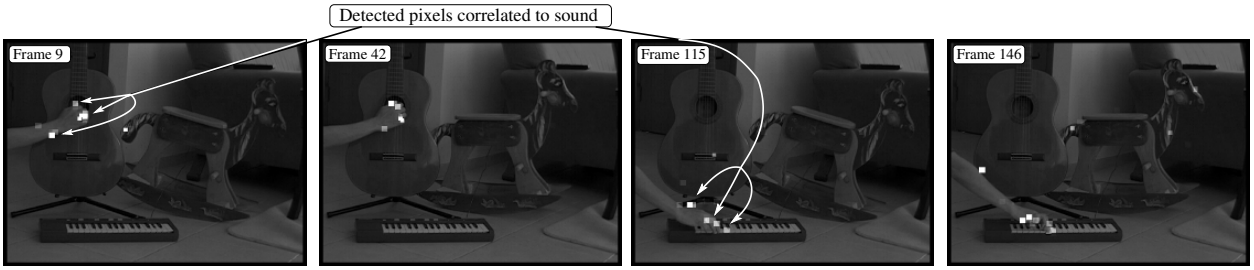

Figure 7. The algorithm results, when run on Movie #1. For visualization, we overlayed the detected energy distribution with the corresponding sample raw frames. Localization concentrates on the playing fingers, which dynamically move from the guitar to the synthesizer. Sporadic detections exist in other areas, usually with much lower energies. **Movie results are linked via http://www.ee.technion.ac.il/∼yoav/AudioVisual.html** .

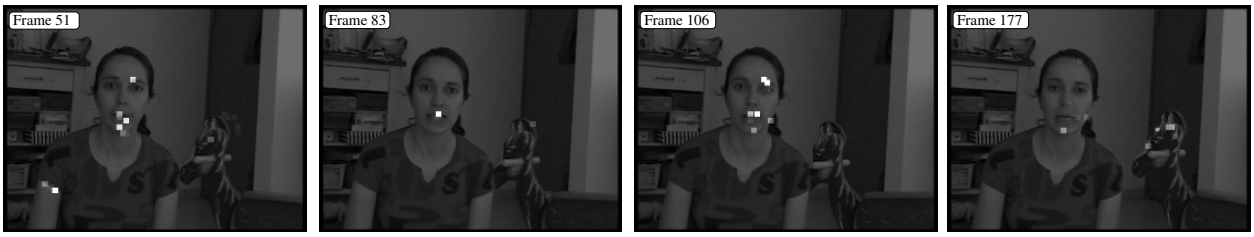

Figure 8. Sample frames resulting from the algorithm, when run on Movie #2. The visualization is as described in Fig. 7. Localization in the mouth area is consistent. Sporadic detections exist in other areas, usually with much lower energies.

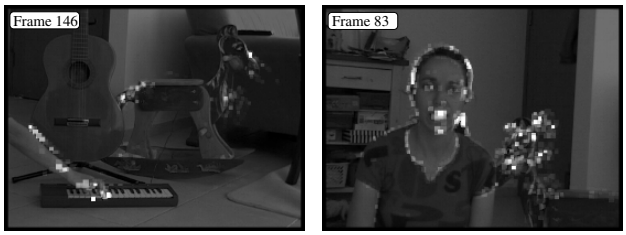

Figure 9. Typical results of using ℓ^2 as a criterion. Compared to the corresponding frames shown in Fig. 7 and 8, the detected energy is much more spread, particularly in non-relevant areas (see the wrong detection of the horse on the right frame).

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

Table 1. The localization criterion L_c obtained in the experiments. The reported numbers are the mean and standard deviation of the measurements. The ℓ^1-norm yields sharp localization, much better than that resulting from ℓ^2.

To judge the results, we compare our algorithm to the performance obtained by using the ℓ^2 criterion, as in (10). Typical sample frames are shown in Fig. 9. They suffer from poor localization and detection rate: there are many false-positives (especially detection on the moving horse), and the energy spreads over a large area. Table 1 reports the temporal mean and standard deviation of the empirical localization values L_c, resulting from the use of either the ℓ^1 or ℓ^2-based localization algorithms. These quantitative results indicate that using the ℓ^2-based solution achieves very 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 within the standard deviation of the L_c values reported in Table 1. Moreover, the multiple audio representation using $N_a = 4$ was tested. The performance was very similar to that described in Figs. 7, 8 and Table 1.

7 Discussion

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

We have presented a robust approach for audio-visual dynamic localization, based on a single microphone. It overcomes the lack of sufficient data (ill-posedness) associated with short time intervals. The algorithm exploits the spatial sparsity of audio-visual events. Furthermore, leaning on recent results that show the relation between sparsity and the ℓ^1-norm, we are able to convexize the problem. Our algorithm is parameter-free, and is thus robust to scenario variability. Nevertheless, the principles posed here can be-

come the base for a more elaborate localization approach, that uses temporal consistency as a prior, as done in tracking methods.

It is possible to extend this approach, e.g., by a kernel version for treating nonlinear relations between the modalities [1, 21, 29]. In addition, time-lag between the audio and the video data can be introduced as a variable in the optimization. Based on the speed of sound, this would enable estimation of object distances from the camera. Furthermore, our sparsity-based approach may be helpful in other scientific domains that aim to correlate arrays of measurement vectors (unrelated to sound), such as climatology.

Acknowledgments

Yoav Schechner is a Landau Fellow - supported by the Taub Foundation, and an Alon Fellow. The work was supported by the US-Israel Binational Science Foundation (BSF), and the Ollendorff Minerva Center in the Elect. Eng. Dept. at the Technion. Minerva is funded through the BMBF.

References

[1] F. Bach and M. Jordan, 2002, "Kernel independent component analysis," J. of Mach. Learning Res. **3**, pp. 1-48.

[2] M. J. Beal, N. Jojic, and H. Attias, 2003, "A graphical model for audiovisual object tracking," IEEE Tran. on PAMI, **25**, pp. 828-836.

[3] C. Bregler, and Y. Konig, 1994, "Eigenlips for robust speech recognition," In Proc. IEEE ICASSP, vol. 2, pp. 667-672.

[4] R. Cutler, and L. Davis, 2000, "Look who's talking: speaker detection using video and audio correlation," Proc. IEEE ICME, vol. 3, pp. 1589-1592.

[5] T. De Bie, and B. De Moor, 2003, "On the regularization of canonical correlation analysis," Int. Sympos. ICA and BSS, pp. 785-790.

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

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

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

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

[10] M. Elad, and M. Zibulevsky, 2004, "A probabilistic study of the average performance of the basis pursuit", submitted to the IEEE Trans. on IT.

[11] G. Farnebäck, 1999, "A unified framework for bases, frames, subspace bases, and subspace frames", Proc. Scand. Conf. Image Analysis pp. 341-349.

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

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

[14] J. W. Fisher III, T. Darrell, W. Freeman, and P. Viola, 2001, "Learning joint statistical models for audio-visual fusion and Segregation," Advanced in Neural Inf. Process. Syst. **13**, pp. 772-778.

[15] R. Gribonval, and M. Nielsen, 2003, "Sparse representations in unions of bases," IEEE Trans. IT **49**, pp. 3320-3325.

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

[17] J. Hershey, and J. Movellan, 1999, "Audio-vision: using audio-visual synchrony to locate sound," Advances in Neural Inf. Process. Syst. **12**, pp. 813-819.

[18] E. Kidron, Y. Y. Schechner, and M. Elad, 2005, "Pixels that sound," Tech. Rep. CCIT TR-524, Dep. of Electrical Engineering, Technion.

[19] H. Knutsson, M. Borga, and T. Landelius, 1995, "Learning canonical correlations," Tech. Rep. LiTH-ISY-R-1761, Computer Vision Laboratory, S-581 83 Linköping Univ., Sweden.

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

[21] T. Melzer, M. Reiter, and H. Bischof, 2003, "Appearance models based on kernel canonical correlation analysis," Patt. Rec. **36**, pp. 1961-1971.

[22] D. Murphy, T. H. Andersen, and K. Jensen, 2004, "Conducting audio files via computer vision," Lecture Notes in Computer Science, **2915**, pp. 529-540

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

[24] C. Schauer, and H. M. Gross, 2003, "A computational model of early auditory-visual integration," Proc. Patt. Rec. Sympos., Lecture Notes in Computer Science **2781** pp. 362-369.

[25] M. Slaney, and M. Covell, 2000, "FaceSync: a linear operator for measuring synchronization of video facial images and audio tracks," Advanc. in Neural Inf. Process. Syst. **13**, pp. 814-820.

[26] M. Song, J. Bu, C. Chen, and N. Li, 2004, "Audio-visual based emotion recognition-a new approach," Proc. IEEE CVPR, vol. 2, pp. 1020-1025.

[27] P. Smaragdis, and M. Casey, 2003, "Audio/Visual independent components," Int. Sympos. ICA and BSS, pp. 709-714.

[28] J. Vermaak, M. Gangnet, A. Blake, and P. Perez, 2001, "Sequential Monte Carlo fusion of sound and vision for speaker tracking," Proc. IEEE ICCV, vol. 1, pp. 741-746.

[29] L. Wolf, A. Shashua, 2003, "Learning over Sets using Kernel Principal Angles," J. of Mach. Learning Res. **4**, pp. 913-931.