The ability to localize visual objects that are associated with an audio source and at the same time separate the audio signal is a cornerstone in several audio-visual signal processing applications. Past efforts usually focused on localizing only the visual objects, without audio separation abilities. Besides, they often rely on computationally expensive pre-processing steps to segment images into object regions before applying localization approaches. We aim to address the problem of audio-visual source localization and separation in an unsupervised manner. The proposed approach employs low-rank in order to model the background visual and audio information and sparsity in order to extract the sparsely correlated components between the audio and visual modalities. In particular, this model decomposes each dataset into a sum of two terms: the low-rank matrices capturing the background uncorrelated information, while the sparse correlated components modelling the sound source in visual modality and the associated sound in audio modality. To this end a novel optimization problem, involving the minimization of nuclear norms and matrix $\ell_1$-norms is solved. We evaluated the proposed method in 1) visual localization and audio separation and 2) visual-assisted audio denoising. The experimental results demonstrate the effectiveness of the proposed method.

**Index Terms**— Audiovisual localization, Audio separation, Multi-modal analysis, Low-rank, Sparsity.

### 1. INTRODUCTION

Audio-visual analysis has recently received increased attention from the signal processing and computer vision communities, enabling the development of a wide range of applications such as audio-visual speech recognition [1], audio-visual source separation [2], and multimedia analysis including person identification from audio-visual resources, audio-visual human robot interaction [3], to name but a few.

In this paper, we aim to localize and separate audio-visual objects without limiting the problem on any specific audio-visual sources (e.g., talking faces [2]). In particular, we focus on robustly localizing the image pixels that are associated with an audio source in videos and at the same time separating the audio signal that is associated with the visual object. These pixels should be distinguished from other moving objects and the audio signal should correspond to the sound produced by visual object, even in the presence of interfering sounds or background noise existing, which are unrelated to the desired object.

Existing approaches in audio-visual object localization aim to identify either the pixels [4, 5, 6, 7] or the object [8, 9] in videos that are most correlated to the audio. The pixel-level approaches usually do not contain pre-processing to segment video images, and directly take image pixels as the visual input and output correlated pixels as the result of localization. In [4], Kdiron et al. used Canonical Correlation Analysis (CCA) to find the correlation of audio and video modalities in order to detect moving sounding objects. In [5],
the problem was handled by a simply coincidence-based measure, which evaluates the correlation between the onsets of audio and visual modalities. Casanovas et al. [7] used nonlinear diffusion to capture the pixels whose motion is most consistent with changes of audio energy, and then applied a graph-cut segmentation procedure [6] to keeps pixels remaining in regions. The object-level approach segments video images into visual atoms or regions before applying localization. In [8], the authors oversegmented each video frame into a number of small segments, and then clustered them to form visual objects. The audio-associated visual object was finally identified via CCA. In [9], Li et al. first applied an region tracking algorithm to segment the video into regions. Then a nonlinear transformation was implemented to obtain both the audio and visual codes in a common rank correlation space. Finally, the correlation was evaluated by computing the hammering distance between the generated codes. However, the aforementioned methods are not able to separate the audio signal associated with the visual objects.

Here, distinct from previous methods we propose a novel method for unsupervised audio-visual source localization and separation using low-rank and sparsity. To this end, we assume that the background of the video lies in a low-dimensional subspace while the moving foreground objects that produce sound can be regarded as relatively sparse within the image sequence. Moreover, a time-frequency distribution (e.g., spectrogram) of the audio signal is assumed to be a superposition of a low-rank and a sparse part, corresponding to spectrogram of the background and the foreground audio produced by the moving objects, respectively. Such assumptions are common in background subtraction [10] and monaural audio separation [11]. Therefore, we seek to express visual and audio representations as superpositions of low-rank and sparse parts, where the low-rank parts capture the background uncorrelated information and the sparse parts account for the correlated audio-visual components, revealing the sound source in visual modality and the associated sound in audio modality. An overview of the proposed method is depicted in Figure 1.

To demonstrate the generality of the proposed method and its algorithmic framework, experiments are performed on two application domains, namely 1) visual localization and audio separation and 2) visual-assisted audio denoising.

2. PROPOSED METHODOLOGY

Consider \( V \in \mathbb{R}^{I_1 \times T} \) and \( A \in \mathbb{R}^{I_2 \times T} \) representing the visual and the audio modalities respectively, where \( T \) is the number of frames in the video. In order to localize the visual object that produces sound and separate its associated audio signal we seek to decompose of each matrix into two terms:

\[
V = B_v + P_v \quad A = B_a + P_a, \tag{1}
\]

where \( B_v \in \mathbb{R}^{I_1 \times T} \), and \( B_a \in \mathbb{R}^{I_2 \times T} \) are the low-rank components capturing the information about background images and background sounds, respectively and \( P_v \in \mathbb{R}^{I_1 \times T} \), and \( P_a \in \mathbb{R}^{I_2 \times T} \) are sparse components, accounting for the foreground moving object in images and the correlated part of sounds respectively.

To ensure that \( P_v \) and \( P_a \) are maximally correlated they are further decomposed as following:

\[
P_v = D_v \cdot C \quad P_a = D_a \cdot C, \tag{2}
\]

where dictionary matrices \( D_v \in \mathbb{R}^{I_1 \times K} \), \( D_a \in \mathbb{R}^{I_2 \times K} \) and \( C \in \mathbb{R}^{K \times T} \) represents a common low-dimensional embedding among the two modalities capturing their correlation [12]. The \( K \) denotes the number of correlated components between the visual and audio information.

A natural estimator accounting for the low rank of the \( B_v, B_a \) components and the sparsity of the correlated \( P_v, P_a \) components, is to minimize the rank of \( B_v \), \( B_a \) and the number of non-zero entries of \( P_v \), \( P_a \) measured by the \( \ell_0 \)-norm, e.g. [10, 13]. Since both the rank and \( \ell_0 \)-norm minimization is NP hard [14, 15], we adopted the technique in the robust PCA, which uses the nuclear norm \( \| \cdot \|_* \) and the \( \ell_1 \)-norm to serve as convex envelopes of the rank and \( \ell_0 \)-norm respectively. Therefore, the objective function of our novel algorithm is defined as following:

\[
\mathcal{F}(B_v, B_a, P_v, P_a) = \|B_v\|_* + \|B_a\|_* + \lambda_1\|P_v\|_1 + \lambda_2\|P_a\|_1,
\]

where and \( \lambda_1, \lambda_2 \) are positive parameters to balance the significance of minimizing the sparsity of \( P_v, P_a \) compared to the rank of \( B_v, B_a \).

Furthermore, to smooth the temporal change of the shared matrix \( C \) in sparse components \( P_v, P_a \), we applied a temporal Laplacian regularization \( \text{trace}(C \cdot L \cdot C^T) \) [16], which encodes the sequential relationships in time series data. Thus we formalize the complete constrained optimization problem as following:

\[
\begin{align*}
\text{minimize} & \quad \|B_v\|_* + \|B_a\|_* + \lambda_1\|P_v\|_1 \\
& \quad + \lambda_2\|P_a\|_1 + \lambda_3 \text{trace}(C \cdot L \cdot C^T) \\
\text{subject to} & \quad V = B_v + P_v \cdot C, \quad A = B_a + P_a \cdot C \\
& \quad P_v = D_v \cdot C, \quad P_a = D_a \cdot C.
\end{align*} \tag{4}
\]

Where the unknown matrices are collected in the set \( \mathcal{V} = \{B_v, B_a, P_v, P_a, D_v, D_a, C\} \), \( \lambda_1, \lambda_2, \lambda_3 > 0 \) are positive parameters and the \( L \) is the constructed Laplacian matrix used to smooth the temporal change of the matrix \( C \).

To solve (4), the Alternating Direction Method of Multipliers (ADMM) is applied here. To this end the on the aug-
mented Lagrange function of (4) is formulated as:
\[
\mathcal{L}(\mathcal{V}, \mathcal{M}) = \|B_v\|_s + \|B_a\|_s + \lambda_1\|P_v\|_1 + \\
\lambda_2\|P_a\|_1 + \lambda_3 \text{trace}(C \cdot L \cdot C^T) + \\
\langle Y - V - B_v - D_v \cdot C \rangle + \frac{\mu}{2}\|V - B_v - D_v \cdot C\|_F^2 + \\
\langle Z - A - B_a - D_a \cdot C \rangle + \frac{\mu}{2}\|A - B_a - D_a \cdot C\|_F^2 + \\
\langle G, D_v \cdot C - P_v \rangle + \frac{\mu}{2}\|D_v \cdot C - P_v\|_F^2 + \\
\langle F, D_a \cdot C - P_a \rangle + \frac{\mu}{2}\|D_a \cdot C - P_a\|_F^2
\]

Where primal variables \( \mathcal{V} \doteq \{B_v, B_a, P_v, P_a, D_v, D_a, C\} \) and \( \mathcal{M} \doteq \{Y, Z, G, F\} \) gathers the Lagrange multipliers associated with the four constraints in (4). Besides, the \( \mu > 0 \) is a positive penalty parameter. The ADMM method minimizes the \( \mathcal{L}(\mathcal{V}, \mathcal{M}) \) with respect to each variable in an alternating fashion and then the Lagrange multipliers get updated at each iteration [17]. The procedure is summarized in Algorithm 1.

Within the algorithm the shrinkage operator \( \mathcal{S}_\tau(x) \) is defined as \( \mathcal{S}_\tau(x) = \text{sgn}(x) \max(|x| - \tau, 0) \) [10], and it is applied to each element in matrices. The singular value thresholding (SVT) operator [18] \( \mathcal{D}_\tau(X) = U \Sigma_s(\Sigma) V^* \) and \( X = U \Sigma V^* \) is any singular value decomposition. Having found the matrices \( B_v, B_a, P_v \), and \( P_a \), the nonzero entries in \( P_v \) indicate the location of pixels that correspond to the moving sound object while the associated audio is obtained by applying the inverse STFT on \( P_a \).

3. EXPERIMENTAL EVALUATION

Datasets: The proposed approach is evaluated on 3 videos which have been used in previous studies and one created by ourselves. We use Violin Yanni and Wooden Horse from [8] and Guitar Solo from [9]. The Wooden Horse and Guitar Solo are challenging videos since they contain other moving objects. We also created an additional video Two Speaker where two subjects uttering two different digits from the CUAVE database [19] are merged in the same frame, whereas the audio signal from only of them is kept.

Visual Evaluation: We follow the evaluation framework in [8, 4]. Firstly, we manually segmented the video images into the regions which are correlated (ground truth) and uncorrelated to the audio signal. Then for evaluation purpose we use the F1 measure and the LE term defined in [4], which provides an evaluation from an energy perspective. The energy of the pixels is defined as: \( e(\vec{x}) = |W_v(\vec{x})|^2 \), where \( W_v \) is a resulted image and \( \vec{x} \) is the pixel coordinate. A satisfactory localization is obtained if most of the energy \( e(\vec{x}) \) is concentrated in the same region of the ground truth. The localization criterion is defined as [4]: \( L_e = \sum_{\vec{x} \in D_e} e(\vec{x}) / \sum_{\vec{x}} e(\vec{x}) \times R_1 + R_2 / R_0 \). Where \( R_1 \) is the ground truth, \( R_2 \) is the manually labeled uncorrelated region with audio, and \( R_0 \) stands for the correctly detected region. Besides, the \( D_e \) represents the set of correctly detected pixels: \( D_e = \{ \vec{x} : e(\vec{x}) > 0 \text{ and } \vec{x} \in R_e \} \).

Audio Evaluation: Following the evaluation framework in [11, 20], we examine the separation results by BSS-EVAL metrics [21]. Specifically, the Source to Distortion Ratio (SDR) is often used to represent the overall performance of audio evaluation. We define the Normalized SDR (NSDR), which only measures the improvement of the SDR between the mixture signal \( \hat{s} \) and the resynthesized sound \( \hat{v} \) from \( P_a \). That is [20]: \( \text{NSDR}(\hat{v}, v, \hat{s}) = \text{SDR}(\hat{v}, v) - \text{SDR}(\hat{s}, v) \), where \( \hat{v} \) is the separated audio signal, \( v \) is the original clean sound, and \( \hat{s} \) is the noisy sound.

Experimental Results on Visual Localization and Audio Separation: Qualitative results for visual localisation are presented in Fig. 2 where the sparse component \( P_v \) is shown. It is clear that the proposed algorithm has successfully identified the sound sources in all the test videos. The hands of the keyboardist, violin player and guitarist in the Wooden Horse, Violin Yanni and Guitar Solo videos, respectively, and the
Our method outperforms space CCA and JIVE the sparse CCA algorithm [4] and the JIVE algorithm [22]. For comparison purposes, we have also implemented hand, sparse CCA and JIVE algorithms capture the moving objects as well in all videos.

Quantitative results for all algorithms shown in Table 1. For comparison purposes, we have also implemented the sparse CCA algorithm [4] and the JIVE algorithm [22]. The proposed algorithm outperforms space CCA and JIVE in terms of $F_1$, $L_c$ for all videos. As shown in Fig. 2 the proposed approach localises quite accurately the audio producing region whereas sparse CCA and JIVE produce many false positive detections. In regard to audio separation, the proposed algorithm outperforms sparse CCA and JIVE in terms of SDR in two videos, Violin Yanni and Two Speaker. As for the videos Wooden Horse and Guitar Solo, the sparse CCA obtains high values of SDR since it fails to capture the correlation between two sensory modalities and simply retains most of the original audio as the sparse component.

**Experimental Results on Visually-Assisted Audio Denoising:** In this section we investigate the capabilities of the proposed algorithm in audio denoising with the assistance of visual information. The audio signal in all videos is corrupted with white noise. The signal to noise ratio is 0 dB. In this scenario, the recovered audio sparse component $P_s$ corresponds to the denoised audio signal. Table 2 shows the quantitative results for all methods. The proposed approach outperforms sparse CCA and JIVE in terms of NSDR in all videos except the last one. In the video Two Speaker, the sparse CCA obtains the NSDR value with 0.06 higher than our algorithm, which means they perform equally well. The results of visual localization are very similar to Fig. 2 so they are omitted due to lack of space. Also in this case the proposed method outperforms sparse CCA and JIVE.

### Table 1. Quantitative evaluations of each algorithm in the case of clean audio input.

<table>
<thead>
<tr>
<th>Video name</th>
<th>Criteria</th>
<th>Sparse CCA</th>
<th>JIVE</th>
<th>Our method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wooden Horse</td>
<td>SDR</td>
<td>32.49±2</td>
<td>5.84±2</td>
<td>15.54±7</td>
</tr>
<tr>
<td></td>
<td>$F_1$</td>
<td>0.06±3</td>
<td>0.20±4</td>
<td>0.36±2</td>
</tr>
<tr>
<td></td>
<td>$L_c$</td>
<td>0.36±3</td>
<td>11.54±3</td>
<td>21.50±5</td>
</tr>
<tr>
<td>Violin Yanni</td>
<td>SDR</td>
<td>10.34±2</td>
<td>4.54±2</td>
<td>10.44±4</td>
</tr>
<tr>
<td></td>
<td>$F_1$</td>
<td>0.15±1</td>
<td>0.22±6</td>
<td>0.31±8</td>
</tr>
<tr>
<td></td>
<td>$L_c$</td>
<td>10.59±3</td>
<td>10.59±7</td>
<td>21.59±5</td>
</tr>
<tr>
<td>Guitar Solo</td>
<td>SDR</td>
<td>31.80±2</td>
<td>11.39±2</td>
<td>27.34±4</td>
</tr>
<tr>
<td></td>
<td>$F_1$</td>
<td>0.15±9</td>
<td>0.14±1</td>
<td>0.37±6</td>
</tr>
<tr>
<td></td>
<td>$L_c$</td>
<td>6.29±9</td>
<td>0.01±5</td>
<td>12.53±7</td>
</tr>
<tr>
<td>Two Speaker</td>
<td>SDR</td>
<td>5.41±0</td>
<td>1.03±1</td>
<td>6.23±7</td>
</tr>
<tr>
<td></td>
<td>$F_1$</td>
<td>0.01±1</td>
<td>0.02±9</td>
<td>0.82±4</td>
</tr>
<tr>
<td></td>
<td>$L_c$</td>
<td>14.82±4</td>
<td>11.14±4</td>
<td>10.57±4</td>
</tr>
</tbody>
</table>

### Table 2. Quantitative evaluations of each algorithm in the case of noisy audio input.

<table>
<thead>
<tr>
<th>Video name</th>
<th>Criteria</th>
<th>Sparse CCA</th>
<th>JIVE</th>
<th>Our method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wooden Horse</td>
<td>NSDR</td>
<td>4.30±4</td>
<td>4.54±2</td>
<td>8.54±5</td>
</tr>
<tr>
<td></td>
<td>$F_1$</td>
<td>0.06±3</td>
<td>0.15±2</td>
<td>0.31±6</td>
</tr>
<tr>
<td></td>
<td>$L_c$</td>
<td>0.36±3</td>
<td>12.59±2</td>
<td>22.19±5</td>
</tr>
<tr>
<td>Violin Yanni</td>
<td>NSDR</td>
<td>5.30±1</td>
<td>4.42±7</td>
<td>9.38±6</td>
</tr>
<tr>
<td></td>
<td>$F_1$</td>
<td>0.15±1</td>
<td>0.22±8</td>
<td>0.31±8</td>
</tr>
<tr>
<td></td>
<td>$L_c$</td>
<td>11.59±3</td>
<td>10.59±7</td>
<td>21.59±5</td>
</tr>
<tr>
<td>Guitar Solo</td>
<td>NSDR</td>
<td>5.70±1</td>
<td>2.60±5</td>
<td>14.06±7</td>
</tr>
<tr>
<td></td>
<td>$F_1$</td>
<td>0.15±9</td>
<td>0.14±1</td>
<td>0.37±6</td>
</tr>
<tr>
<td></td>
<td>$L_c$</td>
<td>6.29±9</td>
<td>0.01±5</td>
<td>12.53±7</td>
</tr>
<tr>
<td>Two Speaker</td>
<td>NSDR</td>
<td>1.36±1</td>
<td>0.82±9</td>
<td>1.36±2</td>
</tr>
<tr>
<td></td>
<td>$F_1$</td>
<td>0.01±1</td>
<td>0.02±9</td>
<td>0.82±4</td>
</tr>
<tr>
<td></td>
<td>$L_c$</td>
<td>14.82±4</td>
<td>11.14±4</td>
<td>10.57±4</td>
</tr>
</tbody>
</table>

Fig. 2. Sample frames of the results of each algorithm. These groups of figures are for video Wooden Horse, Violin Yanni and Guitar Solo. Within each group, each row from top to bottom is the original video frames, the manually labeled ground truth, results produced by sparse CCA, by JIVE algorithm and by our algorithm (from the sparse component $P_s$).

**4. CONCLUSION**

In this paper, we proposed a low-rank and sparse model to handle the visual localization and audio separation problem using pixel intensities and audio spectrogram as visual and audio representations. We conducted two set of experiments: (1) visual localisation and audio separation, and (2) visually-assisted denoising. In both cases, the proposed method correctly identifies the sound source and separates the audio in all the test videos and can also successfully denoise the signal.

**5. ACKNOWLEDGEMENTS**

This work has been funded by the European Community Horizon 2020 under grant agreement no. 645094 (SEWA) and no. 688835 (DE-ENIGMA).
6. REFERENCES


