AUDIO-VISUAL OBJECT LOCALIZATION AND SEPARATION USING LOW-RANK AND SPARSITY

Jie Pu[1]*, Yannis Panagakis*[1]*, Stavros Petridis*[1] *and Maja Pantic*[1,2]

[1]Deparment of Computing, Imperial College London, UK
[2]EEMCS, University of Twente, NL

ABSTRACT

The ability to localize visual objects that are associated with an audio source and at the same time seperate the audio signal is a corner stone 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 computational expensive pre-processing steps to segment images pixels 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 ℓ_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.

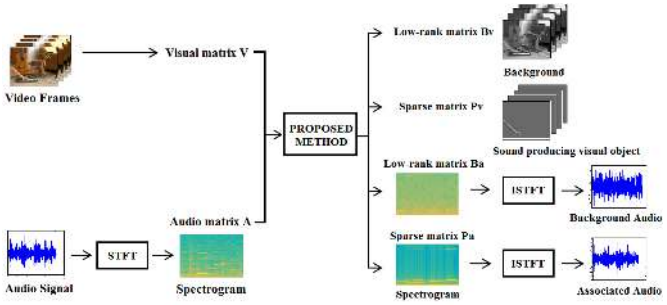


Fig. 1. Overview of the proposed audiovisual source localization and separation method. Matrix $\mathbf{V}$ contains in its columns the vectorized video frames while $\mathbf{A}$ represents the magnitude of the spectrogram, obtained by applying the short-term Fourier transform to the audio signal. The proposed method decomposes $\mathbf{V}$ and $\mathbf{A}$ as superposition of low-rank and sparse parts, where the low-rank matrix $\mathbf{B}_v$ captures the background uncorrelated visual information in video, the low-rank matrix $\mathbf{B}_a$ captures the background audio distraction, while the sparse matrices $\mathbf{P}_v$ and $\mathbf{P}_a$ capture the correlation among visual and acoustic modalities, revealing the location of pixels associated with the sound producing moving visual object as well as its associated spectrogram.

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* approaches segment 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 hamming 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 $\mathbf{V} \in \mathbb{R}^{I_1 \times T}$ and $\mathbf{A} \in \mathbb{R}^{I_2 \times T}$ representing the visual and the audio modalitites respectively, where T is the number of frames in the video. In order to localize the visual object that produces sound and seperate its ascosiated audio signal we seek to decompose of each matrix into two terms:

$$\mathbf{V} = \mathbf{B}_v + \mathbf{P}_v \quad \mathbf{A} = \mathbf{B}_a + \mathbf{P}_a, \tag{1}$$

where $\mathbf{B}_v \in \mathbb{R}^{I_1 \times T}$, and $\mathbf{B}_a \in \mathbb{R}^{I_2 \times T}$ are the low-rank components capturing the information about background images and background sounds, respectively and $\mathbf{P}_v \in \mathbb{R}^{I_1 \times T}$, and $\mathbf{P}_a \in \mathbb{R}^{I_2 \times T}$ are sparse components, accounting for the foreground moving object in images and the correlated part of sounds respectively.

To ensure that $\mathbf{P}_v$ and $\mathbf{P}_a$ are maximally correlated they are further decomposed as following:

$$\mathbf{P}_v = \mathbf{D}_v \cdot \mathbf{C} \quad \mathbf{P}_a = \mathbf{D}_a \cdot \mathbf{C}, \tag{2}$$

where dictionary matrices $\mathbf{D}_v \in \mathbb{R}^{I_1 \times K}$, $\mathbf{D}_a \in \mathbb{R}^{I_2 \times K}$ and $\mathbf{C} \in \mathbb{R}^{K \times T}$ 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 $\mathbf{B}_v$, $\mathbf{B}_a$ components and the sparsity of the correlated $\mathbf{P}_v$, $\mathbf{P}_a$ components, is to minimize the rank of $\mathbf{B}_v$, $\mathbf{B}_a$ and the number of non-zero entries of $\mathbf{P}_v$, $\mathbf{P}_a$ measured by the ℓ_0-norm, e.g. [10, 13]. Since both the rank and ℓ_0-norm minimization is NP hard [14, 15], we adopted the technique in the robust PCA, which uses the nuclear norm $\|.\|_*$ and the ℓ_1-norm to serve as convex envelopes of the rank and ℓ_0-norm respectively. Therefore, the objective function of our novel algorithm is defined as following:

$$\mathcal{F}(\mathbf{B}_v, \mathbf{B}_a, \mathbf{P}_v, \mathbf{P}_a) = \|\mathbf{B}_v\|_* + \|\mathbf{B}_a\|_* + \lambda_1 \|\mathbf{P}_v\|_1 + \lambda_2 \|\mathbf{P}_a\|_1,$$

where and λ_1, λ_2 are positive parameters to balance the significance of minimizing the sparsity of $\mathbf{P}_v$, $\mathbf{P}_a$ compared to the rank of $\mathbf{B}_v$, $\mathbf{B}_a$.

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

$$\begin{aligned} \underset{\mathcal{V}}{\text{minimize}} \quad & \|\mathbf{B}_v\|_* + \|\mathbf{B}_a\|_* + \lambda_1 \|\mathbf{P}_v\|_1 \\ & + \lambda_2 \|\mathbf{P}_a\|_1 + \lambda_3 \, trace(\mathbf{C} \cdot \mathbf{L} \cdot \mathbf{C}^T) \\ \text{subject to} \quad & \mathbf{V} = \mathbf{B}_v + \mathbf{D}_v \cdot \mathbf{C}, \quad \mathbf{A} = \mathbf{B}_a + \mathbf{D}_a \cdot \mathbf{C} \\ & \mathbf{P}_v = \mathbf{D}_v \cdot \mathbf{C}, \quad \mathbf{P}_a = \mathbf{D}_a \cdot \mathbf{C} \end{aligned} \tag{4}.$$

Where the unknown matrices are collected in the set $\mathcal{V} \doteq \{\mathbf{B}_v, \mathbf{B}_a, \mathbf{P}_v, \mathbf{P}_a, \mathbf{D}_v, \mathbf{D}_a, \mathbf{C}\}$, $\lambda_1, \lambda_2, \lambda_3 > 0$ are positive parameters and the $\mathbf{L}$ is the constructed Laplacian matrix used to smooth the temporal change of the matrix $\mathbf{C}$.

To solve (4), the Alternating Direction Method of Multipliers (ADMM) is applied here. To this end the on the aug-

mented Lagrangian function of (4) is formulated as:

$$\mathcal{L}(\mathcal{V},\mathcal{M}) = \|\mathbf{B}_v\|_* + \|\mathbf{B}_a\|_* + \lambda_1\|\mathbf{P}_v\|_1 + \lambda_2\|\mathbf{P}_a\|_1 + \lambda_3\, trace(\mathbf{C}\cdot\mathbf{L}\cdot\mathbf{C}^T) + \langle\mathbf{Y},\mathbf{V}-\mathbf{B}_v-\mathbf{D}_v\cdot\mathbf{C}\rangle + \frac{\mu}{2}\|\mathbf{V}-\mathbf{B}_v-\mathbf{D}_v\cdot\mathbf{C}\|_F^2 + \langle\mathbf{Z},\mathbf{A}-\mathbf{B}_a-\mathbf{D}_a\cdot\mathbf{C}\rangle + \frac{\mu}{2}\|\mathbf{A}-\mathbf{B}_a-\mathbf{D}_a\cdot\mathbf{C}\|_F^2 + \langle\mathbf{G},\mathbf{D}_v\cdot\mathbf{C}-\mathbf{P}_v\rangle + \frac{\mu}{2}\|\mathbf{D}_v\cdot\mathbf{C}-\mathbf{P}_v\|_F^2 + \langle\mathbf{F},\mathbf{D}_a\cdot\mathbf{C}-\mathbf{P}_a\rangle + \frac{\mu}{2}\|\mathbf{D}_a\cdot\mathbf{C}-\mathbf{P}_a\|_F^2$$

Where primal variables $\mathcal{V} \doteq \{\mathbf{B}_v,\mathbf{B}_a,\mathbf{P}_v,\mathbf{P}_a,\mathbf{D}_v,\mathbf{D}_a,\mathbf{C}\}$ and $\mathcal{M} \doteq \{\mathbf{Y},\mathbf{Z},\mathbf{G},\mathbf{F}\}$ 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 $\mathcal{S}_\tau(x) = sgn(x)max(|x|-\tau,0)$ [10], and it is applied to each element in matrices. The singular value thresholding (SVT) operator [18] $\mathcal{D}_\tau(X) = U\mathcal{S}_\tau(\Sigma)V^*$ and $X = U\Sigma V^*$ is any singular value decomposition. Having found the matrices $\mathbf{B}_v,\mathbf{B}_a,\mathbf{P}_v$, and $\mathbf{P}_a$, the nonzero entries in $\mathbf{P}_v$ indicate the location of pixels that correspond to the moving sound object while the associated audio is obtained by applying the inverse STFT on $\mathbf{P}_a$.

Algorithm 1 ADMM solver for (4)

Input: The visual matrix $\mathbf{V}$ and the audio matrix $\mathbf{A}$. Regulariser $\lambda_1,\lambda_2,\lambda_3 > 0$, the Laplacian matrix $\mathbf{L}$.
Initialize: Set $\{\mathbf{B}_v[0],\mathbf{B}_a[0],\mathbf{P}_v[0],\mathbf{P}_a[0],\mathbf{D}_v[0],\mathbf{D}_a[0],\mathbf{C}[0],\mathbf{Y}[0],\mathbf{Z}[0],\mathbf{G}[0],\mathbf{F}[0]\}$ to zero matrices, $\mu > 0$
while not converged **do**
  $\mathbf{B}_v[t+1] \leftarrow \mathcal{D}_{\frac{1}{\mu}}(\mathbf{V}-\mathbf{D}_v[t]\cdot\mathbf{C}[t]+\frac{1}{\mu}\mathbf{Y}[t])$
  $\mathbf{B}_a[t+1] \leftarrow \mathcal{D}_{\frac{1}{\mu}}(\mathbf{A}-\mathbf{D}_a[t]\cdot\mathbf{C}[t]+\frac{1}{\mu}\mathbf{Z}[t])$
  $\mathbf{P}_v[t+1] \leftarrow \mathcal{S}_{\frac{\lambda_1}{\mu}}(\mathbf{D}_v[t]\cdot\mathbf{C}[t]+\frac{1}{\mu}\mathbf{G}[t])$
  $\mathbf{P}_a[t+1] \leftarrow \mathcal{S}_{\frac{\lambda_2}{\mu}}(\mathbf{D}_a[t]\cdot\mathbf{C}[t]+\frac{1}{\mu}\mathbf{F}[t])$
  $\mathbf{D}_v[t+1] \leftarrow \frac{1}{2}(\mathbf{V}-\mathbf{B}_v[t+1]+\frac{1}{\mu}\mathbf{Y}[t]+\mathbf{P}_v[t+1]-\frac{1}{\mu}\mathbf{G}[t])\cdot\mathbf{C}[t]^T\cdot\left(\mathbf{C}[t]\cdot\mathbf{C}[t]^T\right)^{-1}$
  $\mathbf{D}_a[t+1] \leftarrow \frac{1}{2}(\mathbf{A}-\mathbf{B}_a[t+1]+\frac{1}{\mu}\mathbf{Z}[t]+\mathbf{P}_a[t+1]-\frac{1}{\mu}\mathbf{F}[t])\cdot\mathbf{C}[t]^T\cdot\left(\mathbf{C}[t]\cdot\mathbf{C}[t]^T\right)^{-1}$
  $\mathbf{C}[t+1] \leftarrow$ solve the Sylvester equation $\mathbf{MC[t+1]+\ C[t+1]N+\ K=\ 0}$[1]
  $\mathbf{Y}[t+1] \leftarrow \mathbf{Y}[t]+\mu(\mathbf{V}-\mathbf{B}_v[t+1]-\mathbf{D}_v[t+1]\cdot\mathbf{C}[t+1])$
  $\mathbf{Z}[t+1] \leftarrow \mathbf{Z}[t]+\mu(\mathbf{A}-\mathbf{B}_a[t+1]-\mathbf{D}_a[t+1]\cdot\mathbf{C}[t+1])$
  $\mathbf{G}[t+1] \leftarrow \mathbf{G}[t]+\mu(\mathbf{D}_v[t+1]\cdot\mathbf{C}[t+1]-\mathbf{P}_v[t+1])$
  $\mathbf{F}[t+1] \leftarrow \mathbf{F}[t]+\mu(\mathbf{D}_a[t+1]\cdot\mathbf{C}[t+1]-\mathbf{P}_a[t+1])$
  $\mu \leftarrow \min\left(\rho\cdot\mu, 10^{18}\right)$, where ρ is the update factor
  $t \leftarrow t+1$
end while
Output: Background low-rank components $\{\mathbf{B}_v,\mathbf{B}_a\}$, correlated sparse components $\{\mathbf{P}_v,\mathbf{P}_a\}$

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 F_1 measure and the L_c 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]: $L_c = \frac{\sum_{\vec{x}\in\mathcal{D}_c} e(\vec{x})}{\sum_{\vec{x}} e(\vec{x})} \times \frac{R_1+R_2}{R_c}$ Where R_1 is the ground truth, R_2 is the manually labeled uncorrelated region with audio, and R_c stands for the correctly detected region. Besides, the $\mathcal{D}_c$ represents the set of correctly detected pixels: $\mathcal{D}_c \doteq \{\vec{x} : e(\vec{x}) > 0 \ \ and \ \ \vec{x} \in R_1\}$.

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 $\mathbf{P}_a$. 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

Algorithm		Sparse CCA	JIVE	Our method
Video name	criteria			
Wooden Horse	SDR	32.4912	5.6327	15.3204
	F_1	0.0635	0.2040	0.5821
	L_c	3.6232	14.7482	24.2709
Violin Yanni	SDR	7.2470	4.8145	10.4424
	F_1	0.1941	0.2256	0.5138
	L_c	10.9986	10.9917	21.5093
Guitar Solo	SDR	31.3086	11.9821	27.3442
	F_1	0.1509	0.1475	0.3700
	L_c	6.8999	4.3918	12.9377
Two Speaker	SDR	5.4101	1.1031	6.2373
	F_1	0.0111	0.0280	0.4324
	L_c	14.4921	13.1444	193.7176

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

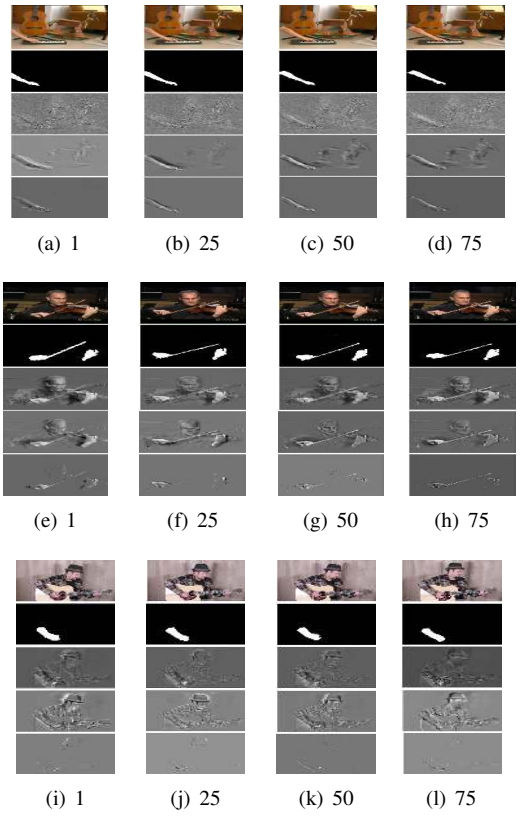
Algorithm		Sparse CCA	JIVE	Our method
Video name	criteria			
Wooden Horse	NSDR	4.3623	4.5420	8.8156
	F_1	0.0635	0.1832	0.5769
	L_c	3.6218	12.8927	24.2161
Violin Yanni	NSDR	5.3031	4.4270	5.8963
	F_1	0.1943	0.2258	0.5165
	L_c	11.4904	10.6450	20.4378
Guitar Solo	NSDR	5.7093	2.6385	14.0807
	F_1	0.1496	0.1478	0.3412
	L_c	6.8872	4.4270	11.8645
Two Speaker	NSDR	1.3641	0.8298	1.3026
	F_1	0.0111	0.0262	0.4156
	L_c	14.4919	11.0700	204.0803

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

mouth of the left subject in the *Two Speaker* video are correctly identified as the correlated sound sources. On the other 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_a 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.



(a) 1 (b) 25 (c) 50 (d) 75

(e) 1 (f) 25 (g) 50 (h) 75

(i) 1 (j) 25 (k) 50 (l) 75

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 $\mathbf{P}_v$).

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

[1] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proceedings of the IEEE*, vol. 91, no. 9, pp. 1306–1326, 2003.

[2] Anna Llagostera Casanovas, Gianluca Monaci, Pierre Vandergheynst, and Rémi Gribonval, "Blind audiovisual source separation based on sparse redundant representations," *IEEE Transactions on Multimedia*, vol. 12, no. 5, pp. 358–371, 2010.

[3] Aggelos K Katsaggelos, Sara Bahaadini, and Rafael Molina, "Audiovisual fusion: Challenges and new approaches," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1635–1653, 2015.

[4] Einat Kidron, Yoav Y Schechner, and Michael Elad, "Pixels that sound," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. IEEE, 2005, vol. 1, pp. 88–95.

[5] Zohar Barzelay and Yoav Y Schechner, "Harmony in motion," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[6] Anna Llagostera Casanovas and Pierre Vandergheynst, "Unsupervised extraction of audio-visual objects," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 2284–2287.

[7] A Llagostera Casanovas and Pierre Vandergheynst, "Nonlinear video diffusion based on audio-video synchrony," *IEEE Trans. on Multimedia*, 2010.

[8] Hamid Izadinia, Imran Saleemi, and Mubarak Shah, "Multimodal analysis for identification and segmentation of moving-sounding objects," *IEEE Transactions on Multimedia*, vol. 15, no. 2, pp. 378–390, 2013.

[9] Kai Li, Jun Ye, and Kien A Hua, "What's making that sound?," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 147–156.

[10] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright, "Robust principal component analysis?," *Journal of the ACM (JACM)*, vol. 58, no. 3, pp. 11, 2011.

[11] Po-Sen Huang, Scott Deeann Chen, Paris Smaragdis, and Mark Hasegawa-Johnson, "Singing-voice separation from monaural recordings using robust principal component analysis," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 57–60.

[12] Gianluca Monaci, Philippe Jost, Pierre Vandergheynst, Boris Mailhe, Sylvain Lesage, and Rémi Gribonval, "Learning multimodal dictionaries," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2272–2283, 2007.

[13] Guangcan Liu and Shuicheng Yan, "Active subspace: Toward scalable low-rank learning," *Neural computation*, vol. 24, no. 12, pp. 3371–3394, 2012.

[14] Lieven Vandenberghe and Stephen Boyd, "Semidefinite programming," *SIAM review*, vol. 38, no. 1, pp. 49–95, 1996.

[15] Balas Kausik Natarajan, "Sparse approximate solutions to linear systems," *SIAM journal on computing*, vol. 24, no. 2, pp. 227–234, 1995.

[16] Sheng Li, Kang Li, and Yun Fu, "Temporal subspace clustering for human motion segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4453–4461.

[17] Yannis Panagakis, Mihalis Nicolaou, Stefanos Zafeiriou, and Maja Pantic, "Robust correlated and individual component analysis," 2015.

[18] Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, no. 4, pp. 1956–1982, 2010.

[19] Eric K. Patterson, Sabri Gurbuz, Zekeriya Tufekci, and John N. Gowdy, "Moving-talker, speaker-independent feature study, and baseline results using the cuave multimodal speech corpus," *EURASIP J. Appl. Signal Process.*, vol. 2002, no. 1, pp. 1189–1201, Jan. 2002.

[20] Zafar Rafii and Bryan Pardo, "A simple music/voice separation method based on the extraction of the repeating musical structure," in *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2011, pp. 221–224.

[21] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte, "Performance measurement in blind audio source separation," *IEEE transactions on audio, speech, and language processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[22] Eric F Lock, Katherine A Hoadley, James Stephen Marron, and Andrew B Nobel, "Joint and individual variation explained (jive) for integrated analysis of multiple data types," *The annals of applied statistics*, vol. 7, no. 1, pp. 523, 2013.