Multimodal Analysis for Identification and Segmentation of Moving-Sounding Objects

Hamid Izadinia, Imran Saleemi, and Mubarak Shah

***Abstract*—In this paper, we propose a novel method that exploits correlation between audio-visual dynamics of a video to segment and localize objects that are the dominant source of audio. Our approach consists of a two-step spatiotemporal segmentation mechanism that relies on velocity and acceleration of moving objects as visual features. Each frame of the video is segmented into regions based on motion and appearance cues using the QuickShift algorithm, which are then clustered over time using K-means, so as to obtain a spatiotemporal video segmentation. The video is represented by motion features computed over individual segments. The Mel-Frequency Cepstral Coefficients (MFCC) of the audio signal, and their first order derivatives are exploited to represent audio. The proposed framework assumes there is a non-trivial correlation between these audio features and the velocity and acceleration of the moving and sounding objects. The canonical correlation analysis (CCA) is utilized to identify the moving objects which are most correlated to the audio signal. In addition to moving-sounding object identification, the same framework is also exploited to solve the problem of audio-video synchronization, and is used to aid interactive segmentation. We evaluate the performance of our proposed method on challenging videos. Our experiments demonstrate significant increase in performance over the state-of-the-art both qualitatively and quantitatively, and validate the feasibility and superiority of our approach.**

***Index Terms*—audio-visual analysis; canonical correlation analysis; video segmentation; audio-visual synchronization;**

I. Introduction

Perceptual organization and grouping in visual data has been a fundamental area of research in multimedia, image processing, and computer vision. The cues indicative of pixels that should be grouped together range from appearance and texture, to shape and motion. The segmentation problem applies to spatial, temporal, as well as spatiotemporal domains, and the efforts focused on these have resulted in numerous approaches. Most of these approaches operate on appearance or motion features in which coherency is sought. For example, in image segmentation, foreground objects can be separated from the background and the image can be partitioned into coherent spatial regions based on features extracted directly from pixel intensities, texture, shape, or edges. Another commonly encountered problem is temporal segmentation of video, e.g., shot boundary detection, where the goal is to group entire contiguous frames that exhibit common features. Segmentation algorithms are also applicable to sequence of images to extract voxels, i.e., regions exhibiting coherency in appearance over time. Several different approaches have been proposed for motion segmentation or detection, and motion-based segmentation. In motion detection, the idea is to separate the moving objects or pixels from a stationary scene, while motion-based segmentation may involve further segmentation of moving objects into coherent moving segments (e.g. the different parts of a human body).

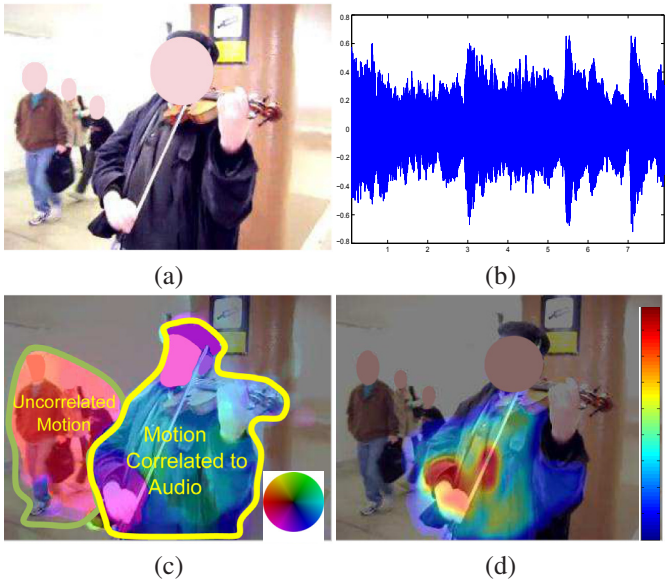

Fig. 1. (a) A sample frame of a video showing a violinist; (b) Audio signal associated with the video, shown as amplitude plot; (c) Optical flow shows the two dominant motions in the scene (flow direction shown by hue and magnitude by brightness); and (d) Localization of moving and sounding objects (the player and the violin in this case) generated by the proposed method, where the probability of correlation to the sound is shown by the colorbar.

In absence of multiple modalities, segmentation techniques incorporate the similarity of features computed in the visual domain to group coherent pixels. On the other hand, multimodal methods attempt to exploit the temporal association of salient events between different modalities. This can be interpreted as localizing the events of one modality with respect to the other. One such area that lends itself to multimodal analysis, especially in computer vision and multimedia, is estimation of association between audio and visual signals. Such correlation is employed for audio source localization based on its analysis with respect to visual events [7], [14], [15], [31]. On the other hand, localization of visual events in auditory mode can be used for audio source separation [6], [1]. The audio-visual analysis is also used for some specific applications such as speaker localization [31], audio-video synchronization [28], and tracking [25].

Consider a scene in which multiple object move simul-

H. Izadinia, I. Saleemi, and M. Shah are with the Computer Vision Lab, Department of Electrical Engineering and Computer Science, University of Central Florida, Orlando, FL, 32816. E-mail: {izadinia, imran, shah}@eecs.ucf.edu

taneously, and some of them emit sounds. This scenario is referred as cocktail party in the literature [6]. An example is illustrated in Fig. 1 (a) where a man plays the violin in a subway station and people walk in the background. Here, an interesting audio-visual analysis would be to find the moving objects whose motion is correlated to the sound of the video. In the example of Fig. 1, the people's motion is not correlated to the audio, while the movement of player and his violin are, and thus should be segmented as moving and sounding objects. The output of such segmentation can be used in higher level recognition and perception systems as it can determine motion of interest in the scene.

In addition to the obvious heterogeneity of audio and video signals, one of the main challenges in audio-visual analysis is their disproportionate dimensionality, i.e., video is a much more complex, high dimensional data compared to audio. There are two extreme approaches to tackle this problem. In the first set of techniques, the dimensionality in each modality is reduced using high level analysis such as face detection [6], [28], [17], [23] and feature point tracking [1]. The other group of approaches, is to analyze the modalities with their original dimensionality intact, and instead using sparse representation of events of the lower dimensional modality based on high dimensional modality under the correlation constraint. The methods [14], [15] assume that the audio sources are spatially localized in the visual mode, and use the sparse representation of low dimensional audio features based on high dimensional visual features.

In this paper, we propose a novel method for segmentation of moving and sounding objects by investigating the maximum correlation between the audio and visual features. The maximum correlation is computed using canonical correlation (CCA) which is a method for finding the maximum correlation between two random variables with different dimensionality. CCA can be considered as an eigensystem problem. For an eigensystem to have a solution, enough samples are needed to estimate the statistics of the signals. Since the correlation is usually analyzed over a small number of frames (i.e., number of samples), we propose to represent audio and visual modalities at a higher level of abstraction. We use the velocity and acceleration of moving objects as visual features which are then grouped together, essentially reducing the dimensionality of the input signal, as well as MFCC and the first derivative of MFCC (MFCC_D) as audio features. We assume that the velocity of objects is correlated to the MFCC features, while their acceleration is correlated to MFCC_D.

We propose a two-step segmentation procedure based on the photometric (color) and dynamic (velocity and acceleration) features of the pixels in spatio-temporal domain, in order to represent the original video in terms of a more local, but high level representation. The velocity and acceleration of each pixel is determined by computing the dense optical flow in each frame. Each component in the final visual feature corresponds to a spatiotemporally local group of pixels, which is intended to represent a semantic object. Once the audio and visual features are computed, the canonical correlation analysis in performed. Each dimension (element) of the *visual* canonical basis vector corresponds to a component or spatiotemporal region in the video, and the value of that element depicts the degree to which that component is correlated to the audio. We therefore select the spatio-temporal segments corresponding to the components with higher value as the moving objects that are most correlated to the observed sound.

The key contribution of our work consists of the following aspects: (i) a two-step spatiotemporal segmentation process for extraction of moving objects, which simultaneously exploits photometric and dynamic features; (ii) a representation of visual and auditory components which lends itself appropriately to correlation analysis; (iii) the use of correlation between velocity and acceleration of moving objects with MFCC and MFCC_D features of the audio signal; (iv) direct utilization of the canonical basis for audio source localization based on the assumption that elements with high values in a canonical basis vector indicate high correlation; and (v) the use of canonical correlation for audio-video synchronization, and interactive video segmentation.

In the following section we briefly review some of the related literature. The proposed method is presented in section III. Then, the experiments and results are discussed in Section IV. The paper is concluded in section V with a summary and concluding remarks.

II. Related Work

Although a large body of work has been proposed in the image, motion, and video segmentation literature [16], as well as audio analysis, we view the proposed framework as part of the much smaller group of methods attempting multimodal analysis. Restricting this review to audio-visual modalities, we observe that research areas in this respect, include vision based source separation [8], music synthesis [21], and reading lips [5], etc.

In terms of localization of sounding objects, multiple methods have attempted to localize moving-sounding objects using arrays of microphones that are calibrated with respect to each other as well as to the cameras. Examples of these techniques include [25], [22], and [24]. However, the problem of sounding object localization using a single microphone (or a single auditory stream) is more challenging. It has been attempted in [20] and [15], although only a single object is localized with the assumption that it is the main contributor to the singular audio stream. The problem of localization of multiple moving-sounding objects has been addressed by Barzelay and Schechner [1].

Specifically, some attention has been focused on the task of localization of visual features, that are associated with audio sources, such that they are distinguished from other uncorrelated moving objects. One class of methods attempting to solve this problem [2] uses stereo triangulation on multiple microphones to localize the sounding object, which is a very strict constraint on practical systems. Another approach motivated by the TREC 2002 monologue detection task, relies on face detection, and employs a mutual information (MI) based synchronization measure between speech and face and lip motion. The method proposed in [7] also exploits mutual information as a synchrony measure, where the multi-

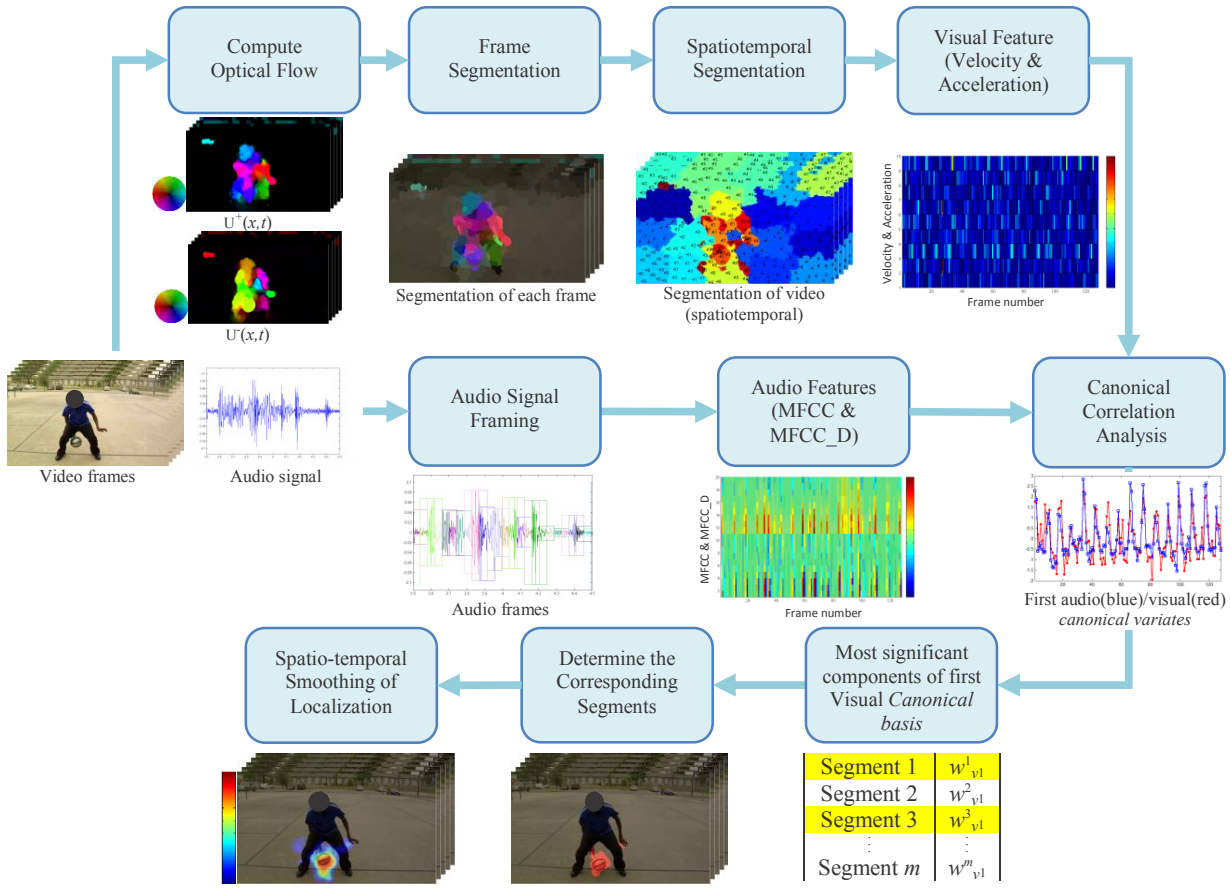

Fig. 2. A flowchart illustrating the proposed method: Computation of visual and audio features is depicted in the first and second rows, respectively. (row 1: L-R) Optical flow between frames t and $t-1$, i.e., $U^{-}(\mathrm{x},t)$, and between t and $t+1$, $U^{+}(\mathrm{x},t)$, where the latter becomes the velocity. Acceleration obtained as difference between the two optical flows; Segmentation preformed in each frame using 3-channel color representations of appearance, velocity and acceleration; K-means used to cluster similar segments over all frames; Mean velocity and acceleration of each spatiotemporal regions represents the entire video. (row 2: L-R) Audio signal framed with half overlap; per frame computation of MFCC and its first derivative (MFCC_D); CCA performed between visual and audio representations. (row 3: R-L) Sorting of components of first visual canonical basis; identification of corresponding spatiotemporal segments as likely sources of audio; and spatiotemporal smoothing of localization mask by Gaussian convolution.

dimensional signals obtained from images of video, and periodograms of the audio signal are adaptively projected to 1-D, such that the Mutual Information (MI) of the representations of each mode is simultaneously optimized to infer the association. In [1], visual interest points are tracked through the video to obtain trajectories. Visual and audio 'onsets' are then computed and correspondences between them are estimated to find correlation between trajectories of the moving objects and sound. Friedland et al [9], have also proposed an interesting framework for localization of actors contributing to the audio signal at a given time instance. Using recordings from a single camera and microphone pair, their proposed method estimates acoustic as well as visual models to determine the number of speakers and estimates "who spoke when". Afterwards, the visual models are used to infer the location of the speakers in the video.

Correlation between signals from distinct modalities or sensors is an obvious technique for analyzing such signals, and one of the most popular multimodal correlation analysis technique frequently used in the literature is Canonical correlation analysis (CCA). CCA has recently been used in as diverse applications as image set matching [30], action recognition [29], speaker identification [27], and camera correlation in multiple camera scenarios [19]. But, to the best of our knowledge, it has not been directly used for audio source localization. In [18], the problem of visual localization of non-stationary sound sources has been formulated as estimation of optimal visual trajectories that are most representative of the motion of sound source in a spatiotemporal volume. The method of [15] does utilize CCA, but to model the association between audio-visual features as a complete correlation between transformed visual and audio features (i.e. $\mathbf{v}w_v = \mathbf{a}$ where $\mathbf{v}$ and $\mathbf{a}$ are visual and audio feature vectors and w_v is a linear transformation for the visual features). In [15], wavelet coefficients of *pixel-level* frame difference are used as the visual feature. However, as the number of visual features is much higher than the number of frames, CCA cannot directly be used to estimate correlation between the audio and visual features. Therefore, the method assumes the sparsity of auditory features in visual modality to investigate a sparse solution for w_v via l_1-norm optimization (i.e. $\min \|w_v\|_1$ subject to $\mathbf{v}w_v = \mathbf{a}$). Our proposed method is unique in that although it exploits CCA for audio-visual correlation, the video representation is based on motion information extracted from spatiotemporally local

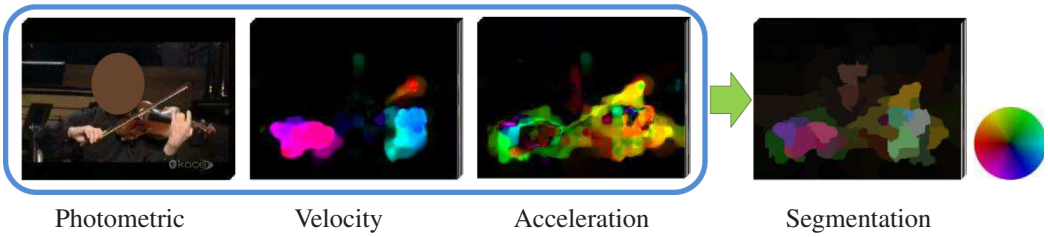

Fig. 3. Segmentation of a frame by using pixel colors, and motion relative to preceding and succeeding frames. Three channel color representation of *photometric*, *velocity* and *acceleration* features, shown for frame 41 of "violin_yanni", as per the color wheel on the right, where color determines the flow direction, and brightness depicts the magnitude. Segmentation result for the frame shown on the right.

regions, as opposed to pixel-level representations. Moreover, in addition to conventionally employed motion and MFCC features, we conjecture the presence of non-trivial association between their rates of change, i.e., first order derivatives.

Another interesting technique for exploiting the relationship between audio and video modalities in multimedia was proposed by Jiang et al [13]. This method proposes to solve the problem of audio-visual concept representation by learning joint codebooks called audio-visual atoms, to improve concept detection. One of the main differences between this approach and the proposed work is that cooccurrence of audio and visual features is not enough to actually localize objects that produce sound. Moreover, unlike learning of joint audio-visual vocabularies, we perform moving-sounding object segmentation instantaneously without training. In other words, methods like [13] require a significant number of training examples to learn models where the proposed work performs similar tasks without any learning.

In [11], Hong et al proposed a framework for dynamic captioning, whereby the system localizes the actor speaking at any given time, and displays the corresponding speech next to the actor to increase video accessability for the hearing impaired. Although this work addresses an interesting and useful application, it has a few limitations in the context of generalizable moving-sounding object localization. Specifically, it does not actually perform video segmentation; assumes availability of actor name and script, thus temporally localizing interesting objects; requires face detection, recognition, and tracking; and is not ideally suited to general, unconstrained videos, as opposed to movie dialog scenes.

Our proposed method on the other hand, is general and self-contained, performs well in unconstrained user uploaded videos, and does not require any other high level computer vision methods. As shown in the test videos (section IV), the moving object could be the face of a talking person, the hand of a player playing an instrument or a moving ball. The proposed framework, detailing feature computation, video representation, as well as correlation, audio source localization and segmentation, is presented next.

III. Proposed Framework

The goal of the proposed work is to detect and segment moving objects observed in a video, which are most correlated to the corresponding audio. Our framework consists of three main steps, (i) extraction of audio and visual features, (ii) canonical correlation analysis over audio-visual features to infer the association between spatiotemporal regions and audio, and (iii) determination of the moving objects in the scene which are correlated to audio features. The overview of the proposed method is presented in Fig. 2, and the details are presented in the following subsections.

A. Feature Extraction and Video Representation

Since the purpose of the proposed method is to accomplish the twofold task of object detection and segmentation, and identification of objects that are highly correlated to audio, the video must be represented as a collection of *local* features, so as to aid estimation of correlation surfaces for *individual* regions or objects in the video. In other words, contrary to conventional representations of motion in audio-visual analysis, i.e., global features like aggregate motion in a frame (histograms of frame difference or optical flow magnitude) or visual disturbance at the frame level, are not useful for our purpose. We therefore, propose to represent the entire video as a collection of photometric and motion features computed on spatiotemporally segmented regions. We now describe our approach for finding such regions.

Fig. 4. The effect of photometric feature in spatiotemporal segmentation presented for two sample frames in each row. In each row, the first column (a) shows the sample frames overlaid by color-coded velocity; (b) shows the result of spatiotemporal segmentation using photometric, velocity and acceleration; while (c) is the result of spatiotemporal segmentation using only velocity and acceleration. The results in (b) are clearly better than those in (c).

Spatiotemporal Video Segmentation: In order to find the moving objects in the video, both static and dynamic features are employed. For static features we use the pixel color while our dynamic features are based on the velocity and acceleration of the pixels which are computed from optical flow. The

selection of motion features is based on the assumption that the velocity of a moving object is correlated to the Mel-Frequency Cepstral Coefficients (MFCC) audio features, whereas the first derivative of velocity (acceleration) is correlated to the first derivative of MFCC features. This is an important, but reasonable assumption since the main idea behind the problem under consideration is that the moving objects in the real world often emit sounds, and moreover, the change in motion often manifests itself as a corresponding change in the observed auditory signal. A simple example of this, is that of a basketball undergoing the dribbling action, where the change in direction of motion of the ball is correlated to the emitted sound.

To formally describe the process of video segmentation, let $\mathrm{U}^{+}(\mathrm{x},t)$ represent the optical flow vector (u,v) at pixel location $\mathrm{x} = (x,y)$, at time t, which is computed between frame F_t and F_{t+1}. Similarly, $\mathrm{U}^{-}(\mathrm{x},t)$ represents the optical flow vector from frame F_t to F_{t-1}. We define the velocity and acceleration vectors as,

$$vel = \mathrm{U}^{+}(\mathrm{x},t), \quad (1)$$

$$acl = \mathrm{U}^{+}(\mathrm{x},t) - (-\mathrm{U}^{-}(\mathrm{x},t)). \quad (2)$$

In order to incorporate the photometric and dynamic features into an efficient representation, the pixels belonging to the same object should be considered as one entity. The conventional segmentation methods for clustering the pixels in the spatio-temporal domain however, impose two problems. First, processing all of the pixels in a whole video is computationally expensive. Second, the regions obtained by the segmentation of all pixels are too noisy. We propose a two-step segmentation process to alleviate these problems. In the first step, we generate an initial segmentation of each frame based on its photometric and dynamic features. Each small segment is then represented as a vector of mean feature values computed over all pixels in the segment. In the second step, the regions or segments computed for all frames are merged into spatiotemporal volumes (or worms) by clustering their feature vector representations, after which each pixel in the spatio-temporal domain then belongs to a single cluster or worm.

The process begins by estimation of optical flow in each frame, followed by velocity and acceleration computation, using equations 1 and 2. In order to map the velocity and acceleration to a similar space as the photometric features (pixel colors) for distance computation and improved visualization, they are converted to the polar representation, and color coded. Specifically, given horizontal and vertical components of velocity vector at every pixel, as (u,v), we first compute the polar representation of the vector, and obtain the hue and intensity (value in HSV) representation of the 2-D vector as the velocity direction and magnitude respectively, where the saturation value is constant. The HSV values corresponding to the acceleration at a given pixel are computed in a similar manner. Notice that while direction is obviously bounded $([-\pi,\pi])$, the magnitude needs to be normalized over the entire video. Each pixel in a single frame is then represented as an 11 dimensional vector, corresponding to the (x,y) location of the pixel, and the 3 color channels (RGB) for each of the photometric, velocity, and acceleration features. The Quick Shift method [32] is then employed to compute the image-motion segmentation in each frame, resulting in a large number of small regions. This step is illustrated visually in Fig. 3, and the influence of photometric features, is evident from Fig. 4.

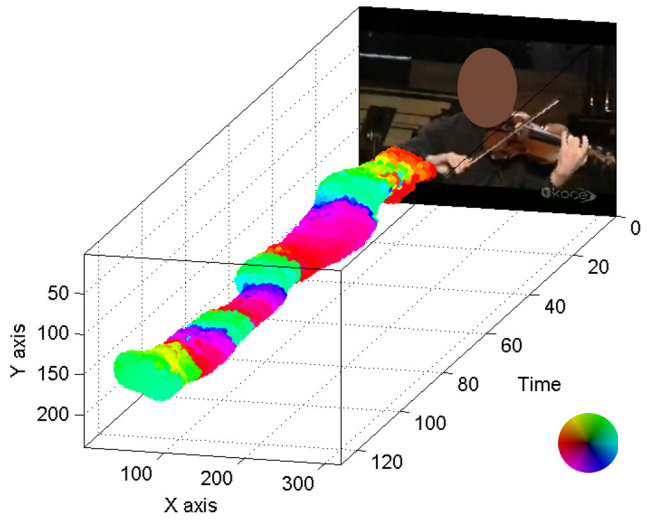

Fig. 5. Example of a typical spatiotemporal volume or worm in a short video clip obtained as a result of our video segmentation process. The optical flow corresponding to each pixel at each frame is visualized by showing flow direction by hue, and magnitude by brightness as per the color wheel. Notice that contiguous, locally smooth flow appears as a worm in the spatiotemporal space.

The second step in the proposed process is the merging of these small segments computed in individual frames, into larger spatiotemporal volumes via clustering. We represent each region of a frame by an 11 dimensional vector, $p = (\mu_{\mathrm{x}}, \mu_c, \mu_{vel}, \mu_{acl})$, where μ_{x} is the 2-D spatial centroid of the region, and each of μ_c, μ_{vel}, and μ_{acl} are 3-D vectors representing mean color, velocity and acceleration of the region. The K-means algorithm with a predefined number of clusters is then used to merge these regions into larger spatiotemporal segments over the entire video, which is the final result of the video segmentation process. An example of such a spatiotemporal segment is visually illustrated in Fig. 5.

Video Representation: Once the segmentation is complete, each pixel in the whole video will be assigned to a single cluster. In order to compute the visual features that will finally represent such a cluster, the average magnitude of velocity and acceleration for all the pixels belonging to the cluster c at frame t is computed as,

$$vel_t^c = \frac{\sum_{p_i \in c} vel(p_i)}{|c|}, \qquad acl_t^c = \frac{\sum_{p_i \in c} acl(p_i)}{|c|}, \quad (3)$$

where $|c|$ denotes the number of pixels belonging to cluster c at frame t. $vel(p_i)$ and $acl(p_i)$ refer to the magnitude of velocity and acceleration for pixel p_i. In order to ignore the clusters that correspond to pixels with nominal motion, we compute the standard deviation of the velocity and acceleration magnitude for each cluster, and sort them in descending order. We then select the top m_1 clusters for velocity, and the top m_2 clusters for acceleration. These features are then concatenated into an

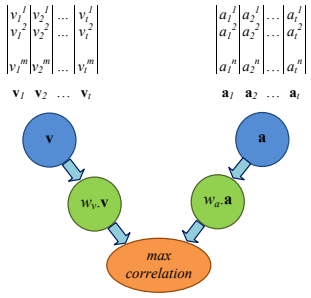

Fig. 6. Canonical Correlation Analysis finds the linear relationships between two $\boldsymbol{m}$ and $\boldsymbol{n}$ dimensional random variables $\mathbf{v}$ and $\mathbf{a}$ using the pairs of transformations $\mathbf{w_v}$ and $\mathbf{w_a}$ called canonical basis.

$m = m_1 + m_2$ dimensional vector for each frame, which then becomes the final visual feature, $\mathbf{v}$, for the video. The feature, $\mathbf{v}$ for a given video, is essentially an $m \times t$ matrix, where the columns, v_i, correspond to each of the t frames in the video, and each element v_i^j, $j \in \{1, \ldots, m\}$, of the i^{th} column is the concatenation of the per-frame mean velocity and acceleration magnitudes of a single spatiotemporal region in the i^{th} frame. The elements of a column corresponding to spatiotemporal regions that are missing in a frame are set to zero.

Audio Representation: It is reasonable to assume that for most realistic videos where the problem is to identify motion corresponding to audio, the auditory signal will be dominated by a single underlying process, i.e., moving object. Therefore, for the representation of the audio signal, we employ the MFCC feature [26], and its derivative, which have often been used in conventional audio processing systems. We employ $\frac{n}{2}$ MFCC coefficients, and the audio signal is then represented as the feature, $\mathbf{a}$, which is an $n \times t$ matrix where the columns, $\mathbf{a}_i$ correspond to the t frames, and each element of a column corresponds to the MFCC and MFCC_D coefficients.

B. Canonical Correlation Analysis

Given the proposed representations of audio and video as the matrices $\mathbf{a}$ and $\mathbf{v}$ respectively, we seek to identify the objects (or regions) in the video that are most correlated to the audio. Assuming that audio is a single entity signal, that is, dominated by a single sound, the problem can be described as finding the dimension (i.e., the spatiotemporal region) in the video feature $\mathbf{v}$, that contributes the most towards maximization of the correlation with $\mathbf{a}$. There are however a few key points to notice in this formulation. First, ordinary correlation will be highly sensitive to the coordinate systems in which $\mathbf{v}$ and $\mathbf{a}$ are described, which are obviously completely different. Second, a simple correlation will not aid our goal of estimating the contribution of key components (rows) of the video feature $\mathbf{v}$, towards the correlation result. We therefore, need a method that not only finds the two optimal bases which project each of the audio and visual features into a common coordinate system, but also simultaneously estimates the corresponding correlations. In other words, we need to perform correlation analysis between $\mathbf{v}$ and $\mathbf{a}$ such that, the correlation matrix between the variables is diagonal, and the diagonal values are maximized.

Fortunately, Canonical Correlation Analysis (CCA) proposed by Hotelling [12], is exactly such a method, which determines the correlation between two multi-dimensional random variables by finding a linear transformation of the first variable that is most correlated to some linear transformation of the second variable. As shown in Fig. 6, the model reveals how well two random variables $\mathbf{v}$ and $\mathbf{a}$ can be transformed to a common source. We use CCA to find pairs of canonical bases $\mathbf{w}_v$ and $\mathbf{w}_a$ in visual and auditory domains respectively, that maximize the correlation between the projections $\mathbf{v}' = \mathbf{w}_v^\top \mathbf{v}$ and $\mathbf{a}' = \mathbf{w}_a^\top \mathbf{a}$ as,

$$\begin{aligned}
\rho &= \max_{\mathbf{w}_v, \mathbf{w}_a} \frac{\mathbb{E}[\mathbf{v}'\mathbf{a}']}{\sqrt{\mathbb{E}[\mathbf{v}'\mathbf{v}'^\top]\mathbb{E}[\mathbf{a}'\mathbf{a}'^\top]}}, \\
&= \max_{\mathbf{w}_v, \mathbf{w}_a} \frac{\mathbb{E}[\mathbf{w}_v^\top \mathbf{v}\mathbf{a}^\top \mathbf{w}_a]}{\sqrt{\mathbb{E}[\mathbf{w}_v^\top \mathbf{v}\mathbf{v}^\top \mathbf{w}_v]\mathbb{E}[\mathbf{w}_a^\top \mathbf{a}\mathbf{a}^\top \mathbf{w}_a]}}, \\
&= \max_{\mathbf{w}_v, \mathbf{w}_a} \frac{\mathbf{w}_v^\top \mathbb{E}[\mathbf{v}\mathbf{a}^\top]\mathbf{w}_a}{\sqrt{\mathbf{w}_v^\top \mathbb{E}[\mathbf{v}\mathbf{v}^\top]\mathbf{w}_v \mathbf{w}_a^\top \mathbb{E}[\mathbf{a}\mathbf{a}^\top]\mathbf{w}_a}}, \\
&= \max_{\mathbf{w}_v, \mathbf{w}_a} \frac{\mathbf{w}_v^\top C_{va} \mathbf{w}_a}{\sqrt{\mathbf{w}_v^\top C_{vv} \mathbf{w}_v \mathbf{w}_a^\top C_{aa} \mathbf{w}_a}},
\end{aligned} \tag{4}$$

wherein $\mathbb{E}[.]$ denotes empirical expectation and ρ is the canonical correlation between random variables $\mathbf{v}$ and $\mathbf{a}$. In this equation, $C_{vv} \in \mathbb{R}^{m \times m}$ and $C_{aa} \in \mathbb{R}^{n \times n}$ are the covariance matrices for $\mathbf{v}$ and $\mathbf{a}$, respectively, while $C_{va} \in \mathbb{R}^{m \times n}$ is the cross-covariance matrix of the vectors $\mathbf{v}$ and $\mathbf{a}$. The covariance matrices are estimated by the total covariance matrix $(\widehat{C})$ defined as,

$$\widehat{C} = \begin{bmatrix} C_{vv} & C_{va} \\ C_{av} & C_{aa} \end{bmatrix} = \mathbb{E}\left[\begin{pmatrix} \mathbf{v} \\ \mathbf{a} \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ \mathbf{a} \end{pmatrix}^\top \right]. \tag{5}$$

Equation 4 has a closed form solution using Lagrange multipliers, and results in a standard eigenproblem as,

$$\begin{cases} C_{vv}^{-1} C_{va} C_{aa}^{-1} C_{av} \mathbf{w}_v = \lambda^2 \mathbf{w}_v \\ C_{aa}^{-1} C_{av} C_{vv}^{-1} C_{va} \mathbf{w}_a = \lambda^2 \mathbf{w}_a, \end{cases} \tag{6}$$

where the corresponding eigenvectors $\mathbf{w}_v$ and $\mathbf{w}_a$ are canonical bases of $\mathbf{v}$ and $\mathbf{a}$, respectively. The eigenvectors corresponding to the largest eigenvalue λ_1^2, are the vectors w_{v^1} and w_{a^1}, which maximize the correlation between the canonical variates, $v_1' = w_{v^1}^\top \mathbf{v}$ and $a_1' = w_{a^1}^\top \mathbf{a}$. For more details, the reader is referred to [10].

C. Localization inference from canonical basis

Once the canonical correlation analysis is performed, the first canonical bases, w_{v^1} and w_{a^1} which lead to maximum correlation, are determined. We only consider the *visual* canonical basis w_{v^1} to localize the moving and sounding objects, since as mentioned earlier, it is reasonable to assume that the entire audio signal is generated by a single underlying process.

Moreover, it is reasonable to assume that in the linear transformation of $\mathbf{v}$ by w_{v^1}, the element with higher value has a larger contribution to the maximum audio-visual correlation, λ_1. We therefore select the elements of w_{v^1}, whose normalized values are more than a predefined threshold (0.5 in our

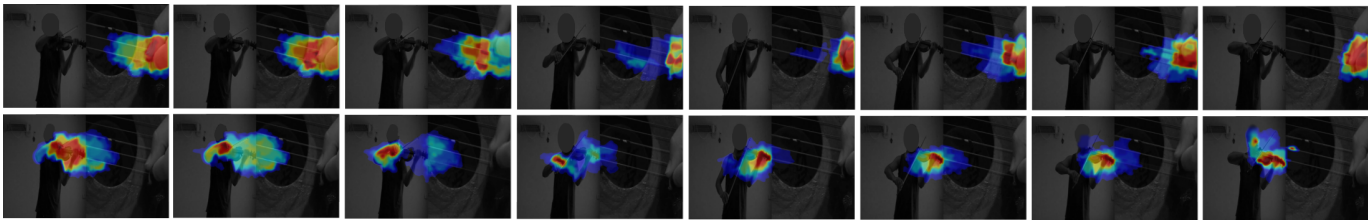

Fig. 7. Localization of two distinct moving objects by exploitation of correlation with corresponding sounds, without explicit sound source separation. Video used in [1] combines two scenes of violin and guitar playing, and the proposed method is able to output probabilities of each pixel belonging to one of the two scenes using the first and second canonical basis after correlation.

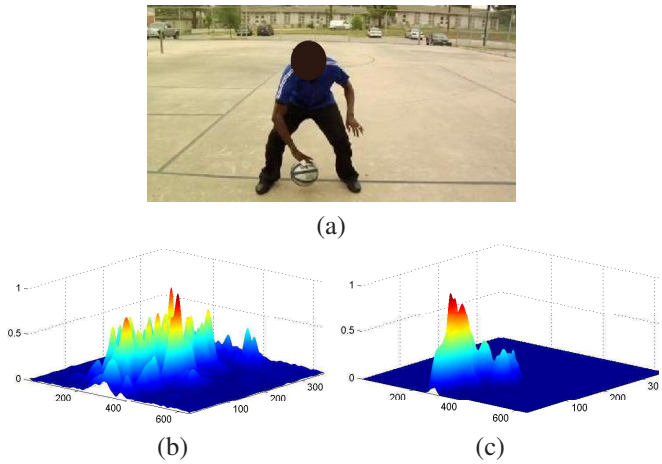

Fig. 8. Probability of localization for moving and sounding object (frame 33 of "basketball"): (a) image frame, (b) baseline method [15], and (c) proposed method. Notice that the probability surface for (c) is much more well defined for the true object, and almost zero otherwise.

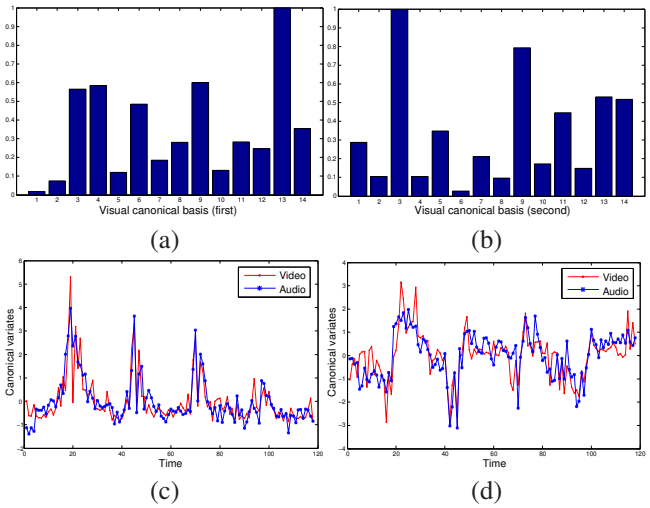

Fig. 9. Quantification of multiple moving-sounding object identification process: (a) normalized values of each element in the first visual canonical basis, w_{v^1}, with the highest value for bin 13, and (b) values in the second canonical basis w_{v^2}, with highest value for bin 3. Each of the 14 bins correspond to one of the spatiotemporal regions identified in the video. (c) and (d) depict the corresponding canonical variates for the first and second basis respectively, where red and blue correspond to visual and audio domains respectively. The similarity of audio and visual variates depicts a high correlation.

experiments). The components (rows) of $\mathbf{v}$ corresponding to the selected elements of w_{v^1}, are chosen as the visual clusters most correlated to the audio. We then set the localization confidence of pixels in those clusters to 1 and the rest to 0. In order to obtain a smooth localization likelihood, we convolve the binary confidences with 2-D and 1-D Gaussian kernels in the spatial and temporal domains, respectively. Consequently the moving objects that are most responsible for producing the audio, are identified as segmented, spatiotemporal regions in the video.

As opposed to a single dominant visual process, e.g., a single object or groups of object moving with high correlation to the single sound source, a more interesting and often prevalent scenario is one where distinct visual processes contribute to the audio signal, without being correlated to, or in sync with, each other. We also observe that the process of audio source localization in the visual domain need not be constrained to using the first canonical variate, v'_1. In other words, while the first canonical bases, w_{v^1}, that leads to maximum correlation, can be used to find the spatiotemporal visual regions that correspond to the audio signal with the highest degree, the second canonical bases, w_{v^2}, may in fact identify regions in the video that correspond to a secondary audio-visual process.

This approach has been tested on videos such as the one shown in Fig. 7, where the two dominant, but out of sync moving-sounding objects are a violin and a guitar, both contributing to the audio signal. Using the first two canonical variates, the proposed method can easily distinguish between the motions corresponding to each of the two objects. The intermediate results of this process have been quantitatively illustrated in Fig. 9 for additional insight.

D. Audio Visual Synchronization

Another straightforward application of the proposed framework of audio and video representation and correlation analysis is the synchronization of audio and video streams. In the previous application of audio source localization, it was assumed that the two streams are synchronized. However, if that is not the case, we can compare the audio and video features in a sliding window fashion. Specifically, assume there is an integer offset of τ^* frames, between the audio and video streams, and we wish to search for the offset in the range $[-\tau_{max}, \tau_{max}]$. We begin by cropping off the first and last τ_{max} frames from the audio stream, and computing the audio feature, $\mathbf{a}$, on the remaining, $t - 2\tau_{max}$ frames. A *set* of video features, $\mathbf{V} = \{\mathbf{v}^k\}$, is then computed for all groups of, $t - 2\tau_{max}$, consecutive frames, of the video, i.e., $k \in [1, 2\tau_{max} + 1]$. Canonical correlation analysis is then per-

formed between the audio feature, $\mathbf{a}$, and *each* video feature, $\mathbf{v}^k$, and the video feature with the maximum correlation is chosen as the correct temporal window corresponding to the cropped audio. Mathematically, this process can be written as,

$$\tau^* = \underset{k}{argmax}\{\lambda_1^{k,\mathbf{a}}\} - (\tau_{max} + 1)\,, \tag{7}$$

where $\lambda_1^{k,\mathbf{a}}$ is the scalar, canonical correlation value, corresponding to the most significant eigenvector, which maximizes the correlation between $\mathbf{v}^k$, and $\mathbf{a}$. In other words, τ^* is an offset (in number of frames), as well as the index of the video clip (among the set of $2\tau_{max}+1$ video clips), such that the canonical correlation between that video clip and the audio feature, $\mathbf{a}$ is maximum, as the cropped audio stream is slided over the entire video.

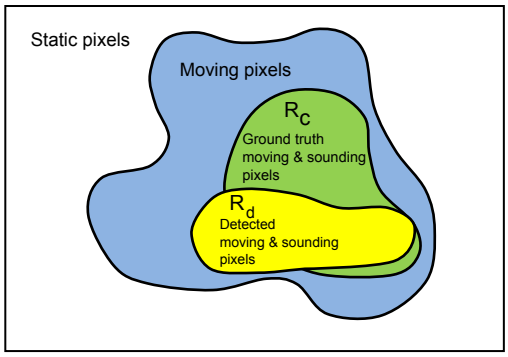

Fig. 10. Quantitative evaluation: All moving pixels occupy the blue region. The ground-truth ($\mathbf{R_c}$) is determined manually as the motion corresponding to sound. Pixels residing in ($\mathbf{R_d}$) are determined by our method, wherein the localization probability is more than a predefined threshold.

IV. Experimental Results

We evaluated our method over several real videos including test videos of [15] and several videos downloaded from YouTube. These videos contain a diversity of scenes representing distinct scenarios of audio-visual processes. The `violin_subway` sequence contains lots of camera motion, illumination changes, and the uncorrelated motion of people in the background. Similarly, `basketball` video contains a moving car, and in `guitar` video one can see a woman moving next to the player. The `news` video has some distracting motion at the bottom of the video. The test video from [15] contains a moving wooden-horse uncorrelated to the audio, and also some synthetic audio noise. Although we do not add any synthetic audio noise to the videos from YouTube, these videos have some natural non-dominant noise. For instance, `basketball` video contains some noise produced by the wind, and in `violin_subway` there is the sound of people walking behind the player. Similarly, in `violin_yanni` the sound of audience clapping can be heard.

We use the original frame rate and resolution of the videos. The audio is sampled at 16KHz and analyzed using a Hamming window with 50% overlap. The length of the audio window is selected such that the audio and video frame rates are synchronized. We use 10 MFCC coefficients along with 10 first derivative of MFCC as audio features. The QuickShift algorithm uses three parameters: the tradeoff between the color and spatial importance (γ), the scale at which the density is estimated (σ), and the maximum distance between pixels in the same region (τ). We used $\gamma = 0.25$, $\sigma = 1$, and $\tau = 15$, for all our experiments. The method of [3] is used for extracting the optical flow. In addition, the number of clusters in K-means is set to 30 and the standard deviation of Gaussian kernel in localization step is set to 5 for spatial and temporal domain.

We quantified the performance of the proposed framework by comparison against the method proposed in [15]. In the experimental setup for the baseline method, the wavelet transform of temporal difference images up to three levels as well as energy of audio signal are used as visual and audio feature vectors, respectively. Following the setting of [15], the videos are analyzed in intervals of 32 frames. The Basis Pursuit algorithm [4] is used for convex approximation of l_1-norm which results in a sparse solution in the visual domain. It should be noted that a spatiotemporal smoothing similar to that in the proposed method, is also applied to the output obtained by the baseline method for providing a fair comparison.

A. Qualitative Evaluation

In this section, the performance of the proposed method is qualitatively evaluated as shown in Fig. 11 which illustrates the localization probability of the proposed method and baseline method overlaid on the sample frames of each video. As shown in Fig. 11, for most of the videos the probability of localization performed using the method of [15] is scattered and also has many false positives which lie outside the true boundary of the audio correlated motion. On the other hand, the proposed method has a strong peak within the correct region and only a few weak false positives, as shown in Fig. 8. More results and videos have been made available online at the authors' website, along with the ground truth annotation.

The superior performance of the proposed method is due to the incorporation of more effective audio and visual features. The method of [15] uses the energy of the signal as the only audio feature which provides limited information about the audio signal. Moreover, the appearance difference features used in [15] usually have significant value for edges which are too noisy. Another representation of the localization probability on a sample frame of `basketball` video is shown in Fig. 8. Since we adopt the dense velocity and acceleration as well as pixel colors in each frame as dynamic feature, a meaningful localization of visual objects is defined based on regions rather than edges.

B. Quantitative Evaluation

In order to evaluate the proposed method quantitatively, we manually segmented the videos into the regions which are correlated and uncorrelated to the audio. This segmentation is treated as the ground truth, which has been made publicly available on our website to encourage comparison with future methods. The performance of the proposed method is compared to the baseline method using precision-recall and hit ratio criteria. The output of the proposed and baseline methods

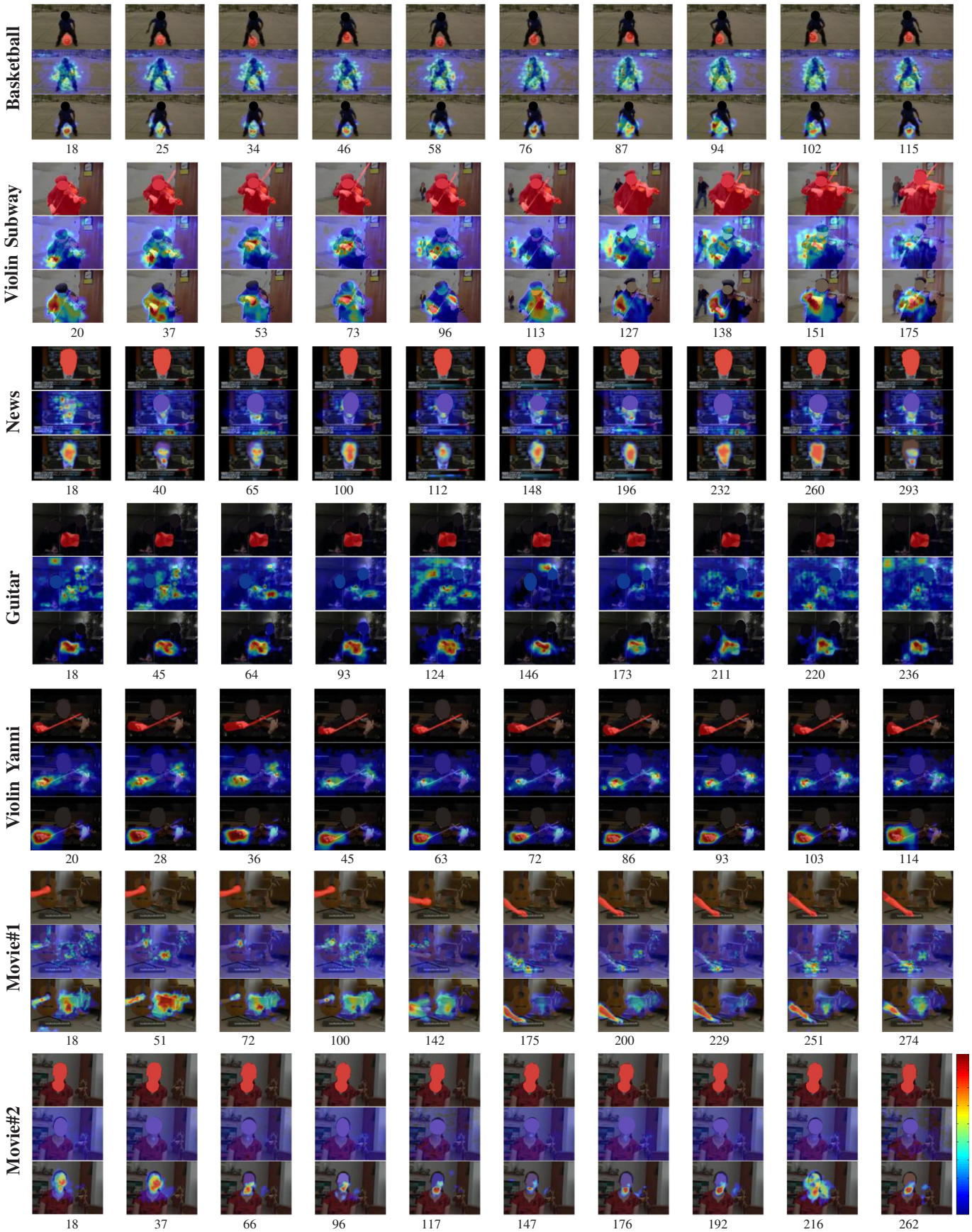

Fig. 11. Localization comparison of the proposed method and baseline method [15] on sample frames of each test video (frame numbers shown at bottom). Movie#1 and Movie#2 are the test videos of [15]. For each video, the first row is the ground truth, second row shows the localization probability produced by the baseline method overlaid on the frame and the third row shows the localization probability obtained by the proposed method. All videos are available at the authors' website.

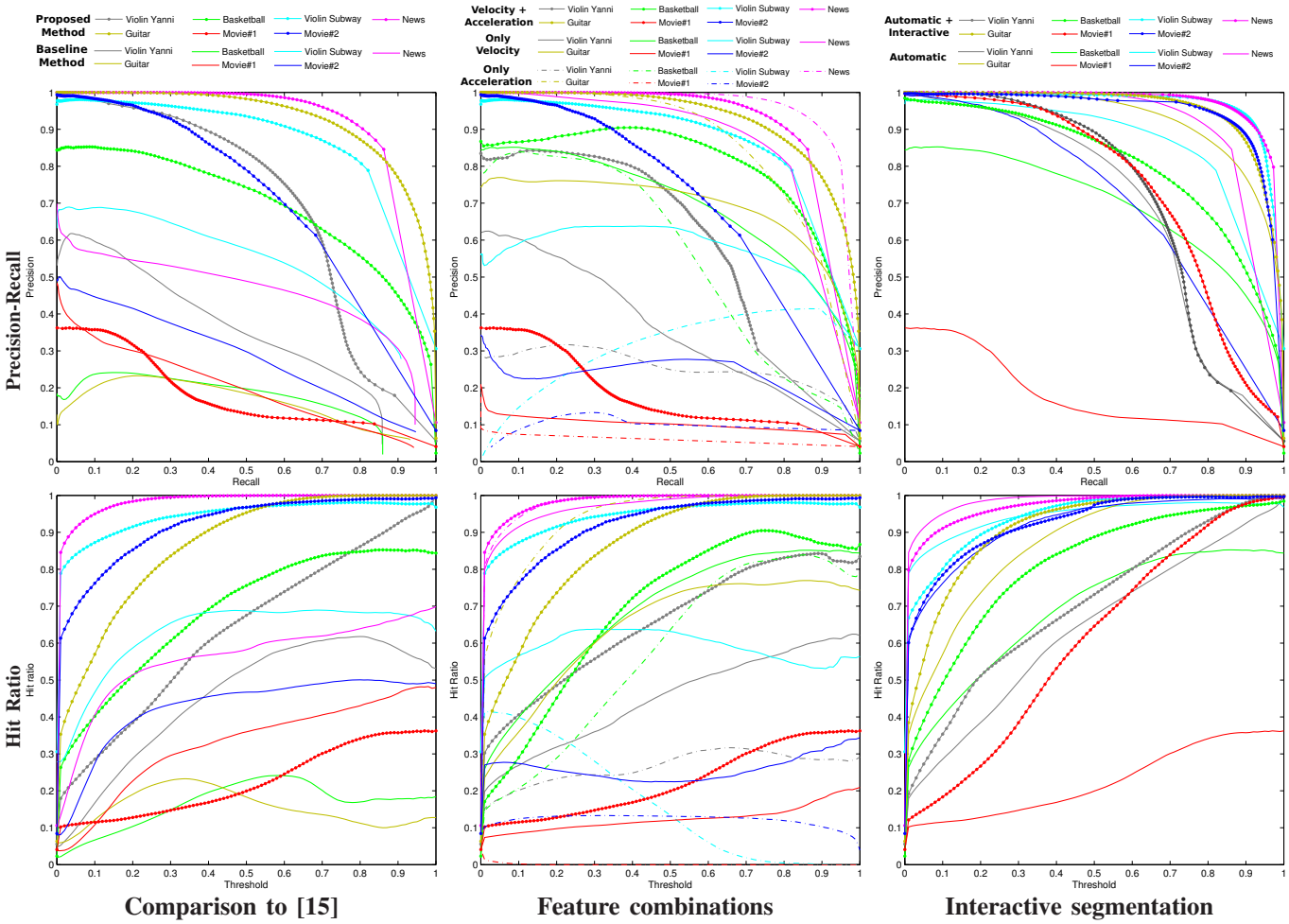

Fig. 12. Localization performance using: (Column 1: top) precision-recall and (bottom) hit ratio, for all test videos by varying threshold from zero to one. The curve of each video is shown by a different color. Results are compared to those obtained using [15]; (Column 2: top) precision-recall and (bottom) hit ratio for all test videos by varying a threshold from zero to one. The proposed method is tested by using different combinations of the proposed motion features as indicated in the legend; and (Column 3: top) precision-recall and (bottom) hit ratio for all test videos by varying a threshold from zero to one. The plots show comparison of audio source segmentation performance using the automated proposed method, and the user input.

for each frame is a surface which shows the probability of each pixel's correlation to the audio. The precision-recall curve is obtained by varying a threshold value from zero to one for each frame. The precision and recall metrics are defined as,

$$precision = \frac{(R_d \cap R_c)}{R_d}; \qquad recall = \frac{(R_d \cap R_c)}{R_c}; \quad (8)$$

where R_c stands for the region correlated to the audio which is annotated manually as ground-truth, and R_d is the localized region that is obtained by the method. The ground-truth of each frame is defined by a contour around the region whose motion is correlated to the audio. This is illustrated in Fig. 10. Since there is a precision-recall curve for each frame, we show the average curve for all frames of a video. Fig. 12(top row, left) shows the comparison between the precision-recall curves of the proposed method and baseline method for all test videos.

In order to explicitly capture the temporal aspect of our performance, we use a second measure called *hit ratio*. A hit occurs in a frame if the precision in that frame is more than 0.5. Hit ratio is defined as the ratio of hits to the number of frames. Fig. 12(bottom row, left) shows the hit ratio of the proposed method compared to the baseline method over all test videos.

As shown in Fig. 12, in all the test videos except `movie#1`, the proposed method has gained higher precision-recall and hit ratio and thus is superior in localizing the moving and sounding objects. The precision-recall and hit ratio of the proposed method obtained in the `movie#1` is very similar to those of the baseline method. This is due to the detection of the wooden horse in the first section of the movie instead of the hand which plays guitar. The underlying reason for detecting wooden-horse is that the motion of the wooden-horse in the first section of the video is largely harmonious with the music played by guitar, while the hand does not have a greatly discernable motion in the captured video.

For evaluating the effect of each of the dynamic features (velocity and acceleration), we run the proposed method using the velocity and acceleration separately. The localization performance of the proposed method with these two different settings are evaluated via precision-recall and hit ratio in Fig. 12(middle column). The results reveal that if the dynamic features are used separately, the performance of

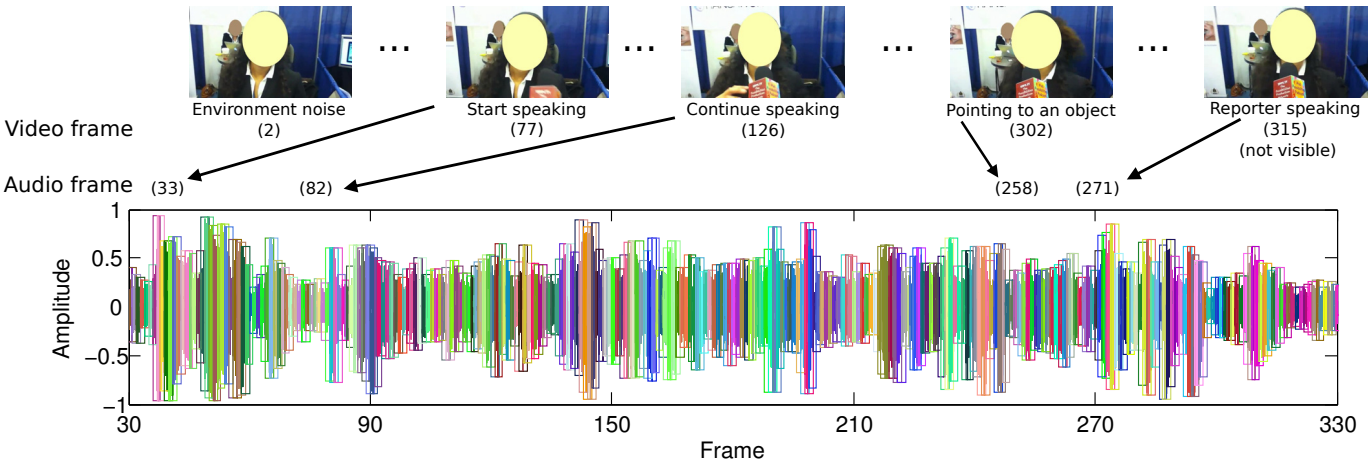

Fig. 13. The original audio (amplitude shown against frames depicted by arbitrarily colored rectangles) and video in this example are clearly out of sync, since the video lags the audio by about 44 frames. As reported in Fig. 14, the offset is correctly identified using the proposed method, thereby synchronizing audio and video.

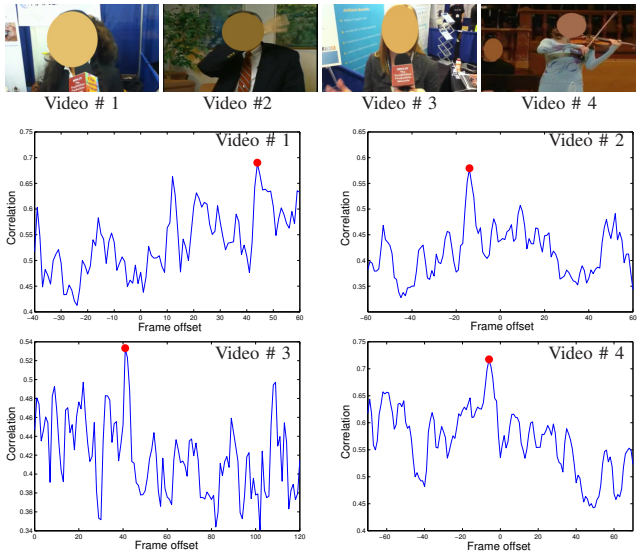

Fig. 14. Audio-video synchronization for 4 video clips. The plots show correlation values (y-axis) using the most significant canonical basis, as the audio and video features are compared in a sliding window fashion from a negative to positive offset (x-axis). Resulting offsets depicted by red circles are: video # 1: +44 frames; video # 2: -14 frames, video # 3: +41 frames, video # 4: -6 frames. Also notice that the correct sync offset is not the only peak in the correlation surface, which demonstrates the complexity of the relationship between the audio and visual domains in these videos.

the proposed method is reduced. Thus, it can be concluded that each of the dynamic features has significant effect on the localization performance of the proposed method. Fig. 12(right column) quantifies the effect of user input on the quality of spatiotemporal segmentation of sounding object which is described in detail later. Finally, Fig. 14 shows results of audio-video synchronization for 4 videos. Details can be found in the caption, and actual examples on the authors' website. The highly multimodal nature of the correlation plots depict the complex nature of the relationship between the audio and video modalities. For the example video shown in Fig. 14(a), an in depth visualization is illustrated in Fig. 13, where the lack of synchrony between temporally salient events, as well as the amount of offset, are evident. Notice that all the videos used in this experiment (Fig. 14) were originally out of sync, and were synchronized manually for quantitative evaluation.

C. Improved Interactive Video Segmentation

Since the spatiotemporal segmentation, on which the visual representation is based, is achieved in a two step process as described earlier, in our framework the first of those steps, i.e., the QuickShift algorithm is allowed to oversegment the image in the spatial domain. Even though the subsequent merging step (K-means) clusters the overly segmented regions belonging to a moving object before computing the final representation, it may still be beneficial to users to allow the opportunity to interactively improve the segmentation for their own purposes. Moreover, although the proposed method performs well in most experimental settings as described in section IV, the sounding object segmentation may not be perfect in some complex scenarios, and a user may be able to guide the process to a better result.

We therefore, also experimented with an interactive method which allows a user to select a few points in an image to indicate foreground and background regions. Typically, the user clicks points only in the keyframes, i.e., every 30 frames of a video. This process is illustrated in Fig. 15. After the user clicks points in the background and foreground regions, the points are joined to form lines. Each overly segmented region's spatial centroid is then compared to the lines using perpendicular Euclidean distance, and the region is labeled as foreground, or 1, if nearer to the foreground indicative lines, and as background, or 0, otherwise. After this process is repeated for a few key frames, spatial and temporal smoothing is performed on the binary masks as described earlier. The two confidence surfaces, i.e., the automatic output of localization by correlation analysis, and interactive estimation, are then combined by simple averaging after normalization.

We also quantified the improvement in the quality of interactive spatiotemporal video segmentation as shown in Fig. 12(right column). It can be observed that the completely automatic proposed method works reasonably well to perform segmentation. However, the user input helps to improve the result by correcting the label of a few regions.

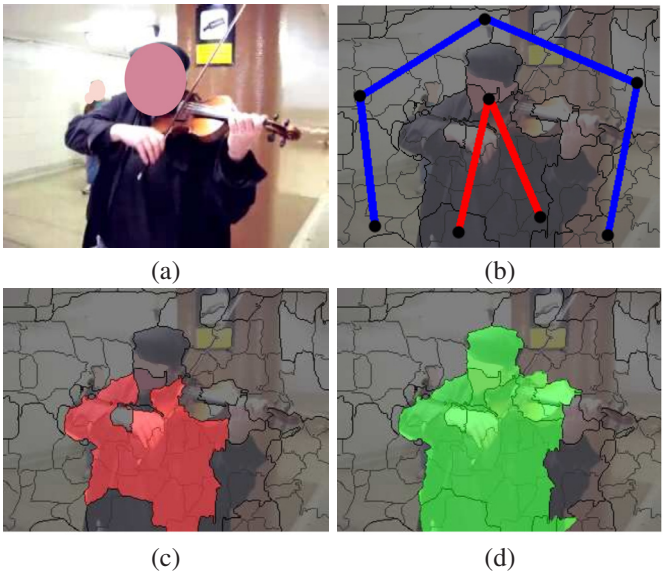

Fig. 15. Interactive video segmentation: (a) original frame; (b) lines formed by joining user clicked points, shown on the image, overlaid by automatically segmented regions. 3 points were clicked to indicate foreground (red), and 5 to indicate background (blue). Bottom row compares the qualitative results of the final segmentation, in (c) using the proposed method, and in (d), by incorporating user input.

V. Conclusion

In conclusion, we have introduced a novel method for detecting the moving and sounding objects via utilization of canonical correlation analysis (CCA). In the proposed method, CCA finds the moving objects whose visual features are most correlated to the audio features. These objects are referred as the moving and sounding objects. To this end, the velocity and acceleration of moving objects are computed based on the dense optical flow of each frame. Then, the moving objects are found via a two-step spatio-temporal segmentation. For the audio features the MFCC and first derivative of MFCC are derived from audio signal. We found the most correlated moving objects based on the assumption that the MFCC and MFCC_D of the audio signal are highly correlated to the velocity and acceleration respectively, of the moving objects that emit sound. The performance of the proposed method is evaluated via experiments on several real videos. The results show that our proposed method can efficiently detect the moving objects that sound, whereas it filters out other dynamics in the scene whose motion is uncorrelated to the audio. Moreover, the same framework is exploited for audio-video synchronization, as well as interactive video segmentation.

VI. Acknowledgement

The research presented in this paper is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20071. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

[1] Z. Barzelay and Y. Schechner. Onsets coincidence for cross-modal analysis. *IEEE Trans. Multimedia*, 12(2):108 –120, 2010.
[2] M. Beal, N. Jojic, and H. Attias. A graphical model for audiovisual object tracking. *IEEE Trans. PAMI*, 25(7):828–836, 2003.
[3] T. Brox and J. Malik. Large displacement optical flow: Descriptor matching in variational motion estimation. *IEEE Trans. PAMI*, 33(3):500 –513, 2011.
[4] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Rev.*, 43:129–159, January 2001.
[5] R. Cutler and L. Davis. Look who's talking: speaker detection using video and audio correlation. In *ICME*, 2000.
[6] T. Darrell, J. Fisher, P. Viola, and W. Freeman. Ausio-visual segmentation and "the cocktail party effect". In *ICMI*, 2000.
[7] J. Fisher and T. Darrell. Speaker association with signal-level audiovisual fusion. *IEEE Trans. Multimedia*, 6(3):406 – 413, 2004.
[8] J. Fisher, T. Darrell, W. Freeman, and P. Viola. Learning Joint Statistical Models for Audio-Visual Fusion and Segregation. In *NIPS*, 2000.
[9] G. Friedland, C. Yeo, and H. Hung. Visual speaker localization aided by acoustic models. In *ACM Multimedia*, 2009.
[10] D. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Comput.*, 16:2639–2664, December 2004.
[11] R. Hong, M. Wang, M. Xu, S. Yan, and T. Chua. Dynamic captioning: video accessibility enhancement for hearing impairment. In *ACM Multimedia*, 2010.
[12] H. Hotelling. Relations between two sets of variates. *Biometrika*, 28:321 –377, 1936.
[13] W. Jiang, C. Cotton, S. Chang, D. Ellis, and A. Loui. Short-term audio-visual atoms for generic video concept classification. In *ACM Multimedia*, 2009.
[14] E. Kidron, Y. Schechner, and M. Elad. Pixels that sound. In *CVPR*, 2005.
[15] E. Kidron, Y. Schechner, and M. Elad. Cross-modal localization via sparsity. *IEEE Trans. Sig. Proc.*, 55(4):1390 –1404, April 2007.
[16] I. Koprinska and S. Carrato. Temporal video segmentation: A survey. In *Sig. Proc.: Img. Comm.*, pages 477–500, 2001.
[17] D. Li, N. Dimitrova, and I. Li, M.and Sethi. Multimedia content processing through cross-modal association. In *ACM Multimedia*, 2003.
[18] Y. Liu and Y. Sato. Visual localization of non-stationary sound sources. In *ACM Multimedia*, 2009.
[19] C. Loy, T. Xiang, and S. Gong. Multi-camera activity correlation analysis. In *CVPR*, 2009.
[20] G. Monaci and P. Vandergheynst. Audiovisual gestalts. In *CVPR Workshop*, 2006.
[21] D. Murphy, T. Andersen, and K. Jensen. Conducting audio files via computer vision. In *Proc. Gesture Workshop*, 2003.
[22] K. Nakadai, K. Hidai, H.G. Okuno, and H. Kitano. Real-time speaker localization and speech separation by audio-visual integration. In *ICRA*, 2002.
[23] H. J. Nock, G. Iyengar, and C. Neti. Assessing face and speech consistency for monologue detection in video. In *ACM Multimedia*, 2002.
[24] A. O'Donovan, R. Duraiswami, and J. Neumann. Microphone arrays as generalized cameras for integrated audio visual processing. In *CVPR*, 2007.
[25] P. Perez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proc. IEEE*, 92(3):495 – 513, March 2004.
[26] L. Rabiner and B-H Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, NJ, 1993.
[27] M.E. Sargin, Y. Yemez, E. Erzin, and A.M. Tekalp. Audiovisual synchronization and fusion using canonical correlation analysis. *IEEE Trans. Multimedia*, 9(7):1396 –1403, Nov. 2007.
[28] M. Slaney and M. Covell. Facesync: A linear operator for measuring synchronization of video facial images and audio tracks. In *NIPS*, 2001.
[29] K. Tae-Kyun and R. Cipolla. Canonical correlation analysis of video volume tensors for action categorization and detection. *IEEE Trans. PAMI*, 31(8):1415 –1428, 2009.
[30] K. Tae-Kyun, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Trans. PAMI*, 29(6):1005 –1018, 2007.
[31] H. Vajaria, S. Sarkar, and R. Kasturi. Exploring co-occurence between speech and body movement for audio-guided video localization. *IEEE Trans. Cir. Sys. Vid. Tech.*, 18(11):1608 –1617, Nov. 2008.
[32] A. Vedaldi and S. Soatto. Quick shift and kernel methods for mode seeking. In *ECCV*, 2008.