EIGEN AND MULTIMODAL ANALYSIS FOR LOCALIZING MOVING SOUNDING OBJECTS

Shreya Khare, Akshay Bhandari and Hema A. Murthy

Indian Institute of Technology Madras, India

ABSTRACT

This paper identifies moving objects in a video that are associated to the corresponding audio, by exploiting the correlation of audio and video features. The proposed technique is based on the correlation of motion features of eigen moving objects with audio mel frequency cepstral coefficients features using canonical correlation analysis. We propose two strategies to detect the eigen moving objects: (i) Per-frame mapped eigen moving object (PFEMO) and (ii) Temporally coherent eigen moving object (TCEMO). While PFEMO segments each frame using superpixel segmentation, TCEMO exploits supervoxel based video segmentation to identify eigen moving objects. Qualitative (mean-opinion score) and quantitative (precision, recall, area under the curve, hit ratio) analysis shows that the performance of the proposed techniques is superior to those of the state-of-the-art methods.

Index Terms— canonical correlation analysis; eigen analysis; multimodal analysis; supervoxels; superpixel

1. INTRODUCTION

Real life events are inherently multi-modal. Humans combine data from multiple sources to form a meaningful and coherent picture of the environment. Multi-modal analysis of audio and video modes has proven to be highly valuable for the task of moving sounding object localization. This is due to its vast applicability in tasks like surveillance, automatic management of video conferences [1], sound-source separation, speech recognition [2], speaker identification [3], etc. The objective of this paper is to identify moving objects in a video which are highly associated to the audio, in a cocktail party scenario [4].

Multi-modal signals are a set of heterogeneous data arrays that exhibit mutual interdependency, as they originate from the same physical phenomena. Thus, there exists a temporal correlation between events in the different modes. Fusion techniques integrate information from different modes to establish relationships between the modes. It has been shown in [5] that combining complementary information from different modes improves the performance of person authentication systems significantly. The varying dimensionality and resolution of the modes make joint analysis challenging. Many methods have attempted to localize moving sounding objects in controlled environments such as conference rooms [6] and lecture rooms [7], using microphone and camera arrays. However, these techniques are not relevant for real-time videos such as those taken by a *single cellphone or surveillance camera*. The challenging task is to track moving sounding objects by fusing the visual signal with a single stereo audio signal.

Existing approaches in the joint audio-visual domain aim to identify the pixels [8–10] or objects in the video [11, 12] that are most correlated to the audio. In [8], pixels associated with the audio were obtained using the correlation between audio and video. This correlation was measured using an estimate of the Mutual Information (MI) between the energy of the audio track and the intensity of a single pixel. In [9], Canonical Correlation Analysis (CCA) was used to correlate audio and video modalities and identify moving sounding pixels. However, this method limits the temporal resolution of coincidence detection. In [10], the task of identifying moving-sounding objects was handled using a coincidence-based measure to evaluate the correlation between the onsets of both audio and video modes. This measure is based on the assumption that if audio-visual events are co-incident, they are related.

Methods dealing with pixel level localization are quite sensitive to visual noise and do not reveal semantic details of the objects present in the video. Synchrony of graph-cut based pixel regions with audio was proposed in [12], to extract audio-visual objects. The non-linear diffusion process proposed in [13] measured coherence between audio and video modes. In [11], each video frame was segmented into smaller segments which were clustered to form spatiotemporal moving objects. Audio sources were identified by correlating Mel Frequency Cepstral Coefficients (MFCC) of the audio using CCA. This approach detects the region of the sound source but is unable to detect the exact shape of the object effectively for all videos.

The objective of this work is to localize moving objects in a video that are correlated to the corresponding audio, using cross-modal analysis. To accomplish this task, motion features of eigen moving objects are correlated to audio MFCC. Eigen analysis is used to identify moving pixels for each frame, which are then used to identify *eigen moving objects*. These eigen moving objects are identified using two methods: Per-frame mapped eigen moving object *PFEMO* and Temporally coherent eigen moving object *TCEMO*. In particular, PFEMO maps the moving pixels to spatial regions using *superpixel* image segmentation on each video frame. These spatial regions are then merged to form spatio temporal clusters using K-means clustering. TCEMO uses *supervoxel* based full video segmentation to obtain eigen moving object. Motion features like mean velocity and acceleration of these clusters are then correlated to MFCC features using CCA. The maximum correlated cluster is identified as the moving sounding object.

The key contributions of this paper are as follows: (i) Eigen analysis has been used as an integral step to identify moving pixels, which are mapped to spatial motion-regions in a video sequence. (ii) The proposed methods, PFEMO and TCEMO, reduce the number of clusters that are correlated with the audio features when compared to [11], thus increasing the performance of CCA and ensuring higher precision and recall. (iii) The proposed approach TCEMO, instead of relying on per frame segmentation of video frames, determines the improved spatiotemporal clusters using supervoxel based video segmentation. This ensures higher hit ratio and considerably improves precision, recall and hit ratio when compared to [11].

The paper is organized as follows. Section 2 describes the proposed framework, followed by experimental evaluation in Section 3.

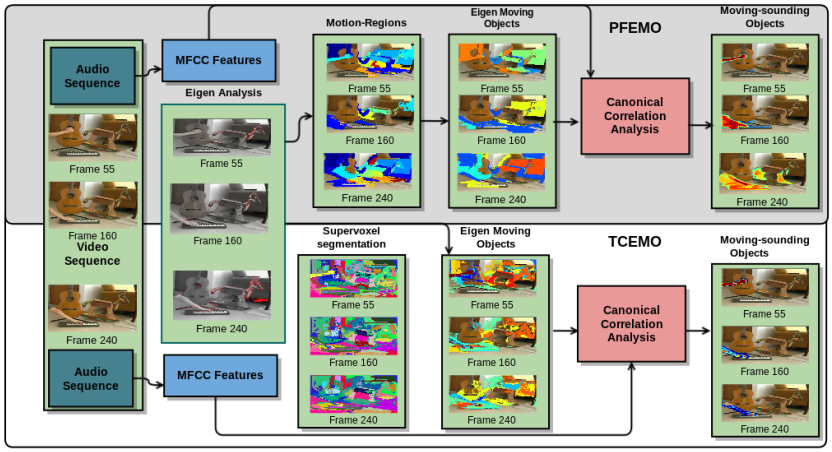

Fig. 1: Overview of the proposed framework. Top and bottom part of the figure show the results of intermediate steps for PFEMO and TCEMO respectively on some sample frames of a video.

The learnings from the effort are summarized in Section 4.

2. PROPOSED FRAMEWORK

A brief overview of the proposed work for localizing moving sounding objects using joint audio-video modes is given in Figure 1. On the whole, the proposed methodology consists of three steps: (i) video modality representation using eigen moving objects (ii) audio modality representation and (iii) object localization using correlation of audio-video modalities.

2.1. Video modality representation using eigen moving object identification

The idea of using eigenspace modeling is to identify pixels that contribute to motion in the video. For an input video $I_{1...T}$ comprising of T frames, an eigenspace model is formed by taking X frames with $X << T$, that captures the stationarity across X images [14]. Since moving objects constantly change their location in the X frames, these objects do not contribute significantly to the eigenspace model as compared to stationary components. Further, I_{i+1} where, $(X+1) \leq i \leq T$, is reconstructed by projecting onto the eigenspace using the top L eigen vectors. Moving pixels are extracted by subtracting the reconstructed frame from the original frame. A threshold is applied to remove noise from the identified moving pixels. Once moving pixels are identified, these pixels are mapped to semantic objects/regions. In particular, two alternative strategies have been developed that map these pixels to temporally coherent objects with fine boundaries. These strategies are described in detail below:

2.1.1. PFEMO-Per frame mapped eigen moving object detection

The identified moving pixels are mapped to regions using superpixel segmentation on each frame. Video features are extracted from these regions and correlated to the audio features. Segmentation of the frames is achieved using a two-pass procedure. In the first pass, each frame is segmented into a large number of regions (superpixels) using Simple Linear Iterative Clustering (SLIC) [15]. In the second pass, *Density based* (Db) clustering [16] is performed to merge superpixels into regions. The regions containing moving pixels above a certain threshold are extracted. These regions are referred to as motion-regions. Figure 1 gives a summary of the proposed method.

These motion-regions across frames are merged into spatiotemporal clusters using K-means. Each motion-region is represented using photometric features, mean velocity and mean acceleration. Velocity and acceleration for each pixel in motion-regions are computed using optical flow. Let $U^+(x,y,t)$ and $U^-(x,y,t)$ denote the optical flow vectors (u,v) between frame F_t-F_{t+1} and F_t-F_{t-1} respectively, for a pixel location (x,y) at time t. Velocity and acceleration of each pixel are computed as

$$\begin{aligned} vel(x,y,t) &= U^+(x,y,t) \\ acc(x,y,t) &= U^+(x,y,t) - (-U^-(x,y,t)) \end{aligned} \tag{1}$$

Each motion-region is then represented by a 11 dimensional feature vector as $R = (\mu_x, \mu_c, \mu_{vel}, \mu_{acc})$ where μ_x is the 2D-mean spatial coordinate and $\mu_c, \mu_{vel}, \mu_{acc}$ are 3D vectors representing mean color, velocity and acceleration of the region, respectively. These motion-regions are then temporally clustered using K-means to form spatiotemporal clusters termed as eigen moving object.

2.1.2. TCEMO -Temporally coherent eigen moving object

We extend the approach of superpixel segmentation used in PFEMO by using supervoxel segmentation. Supervoxels are analogous to superpixels and extend superpixel segmentation by not only clustering the pixels in each image but by segmenting a volumetric stack of images. Besides estimating coherent objects, supervoxels reduce the input complexity when compared to superpixels. A number of supervoxel based video segmentation algorithms have been proposed in [17]. For this work, we have used streaming hierarchical segmentation proposed in [18] which is derived from [19]. The advantage of using this algorithm is that it allows a selection of segmentation levels, preserves the region boundaries, and does not require all voxels in the video to be loaded in memory. Graph-based image segmentation algorithm builds a graph with each pixel as a node, connected with 8 neighbors. For streaming hierarchical video segmentation, this graph is built over the spatiotemporal volume with a 26-neighborhood in 3D space-time for current and previous subsequences (sequence of consecutive frames from the video) of frames.

Eigen moving objects obtained by the above techniques are represented by the average magnitude of velocity and acceleration. A feature vector of dimension $2M \times T$ is obtained by concatenating M clusters each of velocity and acceleration in all the frames. While, for PFEMO, $M = K$ (number of clusters in K-means), for TCEMO, top M clusters with high standard deviation each from velocity and acceleration are chosen.

2.2. Audio modality representation

The audio signal has been represented using MFCC [20] and their derivatives. MFCC are robust features that approximate the human auditory system. The derivatives capture the temporal evolution of the audio signal. The dynamics of video are correlated to the dynamics of audio, as they originate from the same physical phenomena.

2.3. Object localization using correlation of audio-video modalities

Once the audio and video features are extracted, the next step is to identify the regions in the video that are most associated to the audio. To identify the moving sounding objects, we need to discover the hidden correspondence between the two modalities. CCA [21] is a method that determines the correlation between two sets of random variables of different dimension by projecting them on a common coordinate system. For two sets represented as $u \in \mathbb{R}^{d_u}$ and $v \in \mathbb{R}^{d_v}$, CCA effectively projects to subspace $w_u \in \mathbb{R}^{d_u}$ and $w_v \in \mathbb{R}^{d_v}$ such that correlation between the pair of random variables $w_u^T u$ and $w_v^T v$ is maximized as given below:

$$\rho = \max_{w_v, w_u} \frac{\mathbb{E}[w_u^T u v^T w_v]}{\sqrt{\mathbb{E}[w_u^T u u^T w_u]\mathbb{E}[w_v T v v^T w_v]}} \tag{2}$$

where $\mathbb{E}$ denotes the expectatoion and ρ denotes the correlation coeffiecient. The optimal projections are the eigen vectors corresponding to the largest eigen values of the following generalized eigen systems:

$$\begin{aligned} \Sigma_{uv}\Sigma_{vv}^{-1}\Sigma_{vu}w_u &= \lambda_u \Sigma_{uu} w_u \\ \Sigma_{vu}\Sigma_{uu}^{-1}\Sigma_{uv}w_y &= \lambda_v \Sigma_{vv} w_v \end{aligned} \tag{3}$$

where w_u and w_v are the canonical bases of u and v respectively. In equation 3, Σ_{uv} is the cross-covariance matrix of u and v; Σ_{uu} and Σ_{vv} denote the covariance matrix of u and v. The eigen vectors w_{u1} and w_{v1} corresponds to the largest eigen values λ_{u1} and λ_{v1}. Values above a certain threshold in normalized w_{u1} are used to localize the objects corresponding to the most dominant audio source.

Once these objects are localized, a confidence map is created and the localization confidence is set to 1 for the pixels belonging to these objects and is set to 0 for the rest. This confidence map is then convolved with a Gaussian kernel in both spatial and temporal domains to generate a smooth surface. These regions correspond to motion in the video that is most correlated to the audio.

3. EXPERIMENTS AND RESULTS

The proposed methods are evaluated over several real-time test video sequences from [9] and [11]. These videos depict various cluttered environments and scenarios. In *Movie1* video, while the motion of the hand is correlated to the audio, movement of the horse is uncorrelated. The *News* video has uncorrelated movements through highlights, besides the correlated motion of the newsreader's face. In the *Guitar* video, the correlated action is guitar strumming, and the uncorrelated action is the movement of the lady sitting next to the guitarist. In *Violin Yanni (VY)*, the violin bowing is correlated to the violin audio. Moreover, in the audio channel of all the videos, some amount of non-dominant noise is present.

To estimate moving pixels in each frame, a sample size of 6 frames was used for eigen analysis with a threshold of 0.1 for the removal of noise. Once these pixels were obtained, in PFEMO, to obtain segments in each frame, number of superpixels, slic parameter and compactness were set as 25, 10 and 11 respectively. On the other hand, in TCEMO, supervoxels were extracted from hierarchical level 17 with a merging threshold of 10. The audio MFCC features were extracted using Hamming window with 50% overlap. A threshold of 0.5 was used to choose eigen moving objects from CCA. Finally, these objects were localized with a Gaussian kernel of standard deviation 5.

3.1. Qualitative Evaluation

Figure 3 shows the localization probability of the proposed methods overlaid on a few original frames. To accurately assess the method, complete video sequences have been uploaded at the link [1]. It is evident that both the methods have fewer false positives. High localization probability is assigned to regions associated to the audio. In particular, for *Movie-1*, our approach detects only the hand throughout the video. This indicates a significant improvement over the

[1] *https://sites.google.com/site/movingsoundingobjectsresults/home*

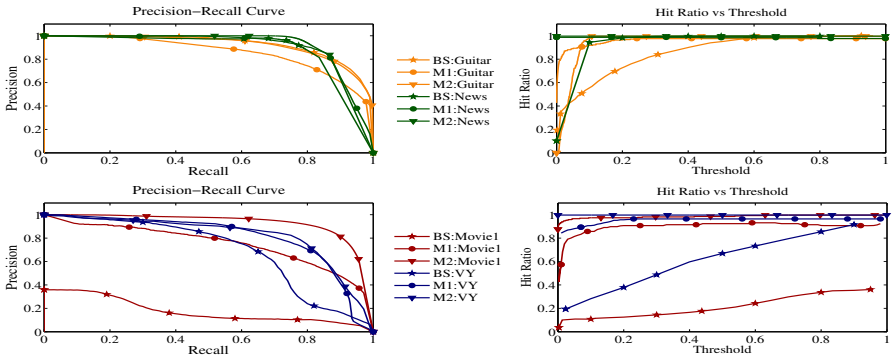

Fig. 2: Quantitative Evaluation: Precision and Hit-Ratio for Baseline [11] (BS), PFEMO(M1) and TCEMO(M2) on 4 videos

baseline method [11], where both the girl and horse are detected. It can also be observed that fine boundaries for the hand are obtained using TCEMO. Similarly, for *Violin-Yanni*, the bowing action is completely detected using both the methods. TCEMO detects the bow stick completely compared to PFEMO. In the *News* video, the newsreader's complete face is detected across all the frames, as compared to the baseline approach. Further, for the *Guitar* video, the guitar is detected with high probability when compared to the baseline method. These inferences are further verified by Mean Opinion Score (MOS) based subjective measurements of the videos generated. These evaluations were carried out with 30 viewers using dou-

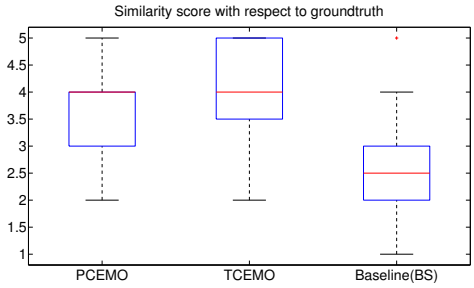

Fig. 4: Boxplot of MOS for PFEMO, TCEMO and BS [11] on 4 videos

ble stimulus method where viewers are asked to rate video clips on a scale from 1(Poor) to 5(Excellent) given the ground-truth video marked by [11] . The boxplot of MOS on similarity to groundtruth is shown in Figure 4. It is clear from Figure 4 that both the proposed systems perform much better than the baseline.

3.2. Quantitative Evaluation

Once the probability of every pixel in all the frames is obtained, it is compared with the ground truth. The performance of each method is evaluated using Precision, Recall and Hit ratio criteria at different threshold values on the localized surface of each frame. A hit occurs in a frame if the precision in that frame is more than 0.5. Hit ratio is

Table 1: Area under the curve in PR and HR curves for baseline(BS) [11], PFEMO and TCEMO

	BS	PFEMO	TCEMO	BS	PFEMO	TCEMO
VY	0.66	0.80	0.81	0.52	0.94	0.99
Movie1	0.17	0.75	0.87	0.22	0.89	0.98
News	0.81	0.90	0.92	0.74	0.98	0.99
Guitar	0.92	0.85	0.93	0.82	0.95	0.96

defined as the ratio of the number of hits to the number of frames. It assists in the evaluation of the temporal coherence of moving sounding objects detected. Figure 2 shows localization performance using precision-recall (PR) and hit ratio (HR) curves for test videos by varying the threshold from 0 to 1. It is evident that higher values are achieved for all the metrics over the baseline method for all the test videos. A significant improvement is obtained by the proposed approaches for *Movie-1* over the baseline method. This is further corroborated by measuring the area under the PR and HR curves for the proposed and the baseline methods as shown in Table 1.

4. CONCLUSION

This paper presents a robust and efficient approach to localization of moving objects that are associated to corresponding audio obtained from a single camera and a single microphone. The proposed methods, namely PFEMO and TCEMO, improve the performance over the current state-of-the-art methods for identifying moving sounding objects. Due to the incorporation of streaming hierarchical video segmentation, TCEMO gives finer and stable boundaries for the moving sounding objects than PFEMO.

The evaluation of our methods clearly indicates that the proposed methods have superior performance when compared to [11]. This can be attributed to two reasons (i) incorporation of eigen analysis removes the extraneous clusters and hence improves CCA. This reduces computations and number of parameters when compared to the method used in [11]. (ii) the proposed segmentation techniques used in PFEMO and TCEMO improve the boundary of the detected objects. This is reflected in the precision, recall and hit ratio for all the test videos. High MOS scores for the proposed systems are indicative of the improved performance.

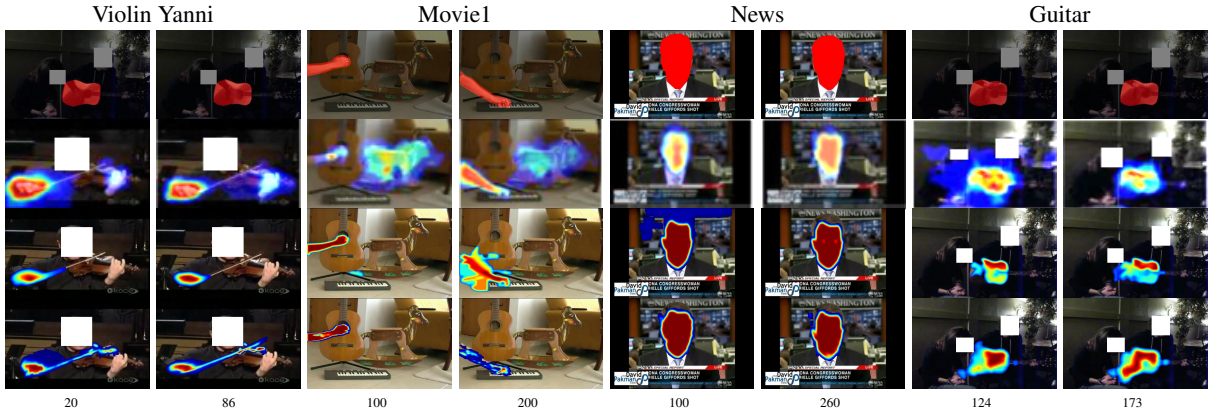

Fig. 3: Qualitative Evaluation:Identification of moving sounding objects on some sample frames for test videos. Results on full video clips can be found at [1]. Top and second rows indicate the ground truth objects in red color and baseline results respectively. Third and fourth rows depict the moving sounding objects identified using PFEMO and TCEMO respectively. Numbers below the fourth row show the frame number in the videos. The different colors correspond to the probability of the moving objects with red corresponding to 1.

5. REFERENCES

[1] K. Otsuka, S. Araki, K. Ishizuka, M. Fujimoto, M. Heinrich, and J. Yamato, "A realtime multimodal system for analyzing group meetings by combining face pose tracking and speaker diarization," in *Proc. ACM ICMI*, 2008, pp. 257–264.

[2] Vahid Asadpour, Mohammad Mehdi Homayounpour, and Farzad Towhidkhah, "Audio–visual speaker identification using dynamic facial movements and utterance phonetic content," *Applied Soft Computing*, vol. 11, no. 2, pp. 2083–2093, 2011.

[3] A. L. Casanovas, G. Monaci, P. Vandergheynst, and R. Gribonval, "Blind audiovisual source separation based on sparse redundant representations," *Multimedia, IEEE Transactions on*, vol. 12, no. 5, pp. 358–371, 2010.

[4] Trevor Darrell, III Fisher, JohnW., and Paul Viola, "Audio-visual segmentation and the cocktail party effect," in *ICMI 2000*, vol. 1948, pp. 32–40. Springer Berlin Heidelberg, 2000.

[5] Samy Bengio, "Multimodal authentication using asynchronous hmms," in *Audio-and Video-Based Biometric Person Authentication*. Springer, 2003, pp. 770–777.

[6] Dmitry N Zotkin, Ramani Duraiswami, and Larry S Davis, "Joint audio-visual tracking using particle filters," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 1, pp. 1154–1164, 2002.

[7] Kai Nickel, Tobias Gehrig, Hazim K. Ekenel, John McDonough, and Rainer Stiefelhagen, "An audio-visual particle filter for speaker tracking on the clear'06 evaluation dataset," 2007, CLEAR'06, pp. 69–80, Springer-Verlag.

[8] John Hershey and Javier Movellan, "Audio-vision: Using audio-visual synchrony to locate sounds," in *Advances in Neural Information Processing Systems 12*. Citeseer, 2000.

[9] Einat Kidron, Yoav Y Schechner, and Michael Elad, "Pixels that sound," in *CVPR 2005*. IEEE, 2005, vol. 1, pp. 88–95.

[10] Zohar Barzelay and Yoav Y Schechner, "Harmony in motion," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

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

[12] A.L. Casanovas and P. Vandergheynst, "Unsupervised extraction of audio-visual objects," in *Acoustics, Speech and Signal Processing (ICASSP), 2011 IEEE International Conference on*, May 2011, pp. 2284–2287.

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

[14] N.M. Oliver, B. Rosario, and A.P. Pentland, "A bayesian computer vision system for modeling human interactions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 831–843, Aug 2000.

[15] Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, pp. 2274–2282, 2012.

[16] Hans-Peter Kriegel, Peer Kröger, Jörg Sander, and Arthur Zimek, "Density-based clustering," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 231–240, 2011.

[17] Chenliang Xu and J.J. Corso, "Evaluation of super-voxel methods for early video processing," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, June 2012, pp. 1202–1209.

[18] Chenliang Xu, Caiming Xiong, and Jason J Corso, "Streaming hierarchical video segmentation," in *Computer Vision–ECCV 2012*, pp. 626–639. Springer, 2012.

[19] Pedro F. Felzenszwalb and Daniel P. Huttenlocher, "Efficient graph-based image segmentation," *Int. J. Comput. Vision*, vol. 59, no. 2, pp. 167–181, Sept. 2004.

[20] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Acoustics, Speech and Signal Processing, IEEE Transactions on*, vol. 28, no. 4, pp. 357–366, Aug 1980.

[21] Harold Hotelling, "Relations between two sets of variates," *Biometrika*, pp. 321–377, 1936.