

Association of Sound to Motion in Video using Perceptual Organization

Abstract

Traditionally, video surveillance has mostly made use of visual data. However, in light of the new behavioral and physiological studies which demonstrated the existence of cross modality effects in human perception, similar cues are being used to develop a surveillance system based on both audio and visual data. Human beings can easily associate a particular sound to an object in the surrounding. Drawing from such studies, we demonstrate a technique by which we can isolate concurrent audio and video events and associate them based on perceptual grouping principles. Simple cues from both audio and video suffice to make this association. By representing audio in the pitch-time domain, we can use image processing algorithms such as line detection to extract elementary audio events. These events are then grouped using Gestalt principles of similarity and proximity into appropriate auditory events. Properties such as time of occurrence and periodicity are easily calculated from these groups. In video, we extract motion and shape periodicities. By comparing all the periodicities in audio and video using a simple index we can easily associate audio to video.

1. Introduction

Recent experiments have shown that human perception does not process visual, sound and smell separately but rather perceives a scene based on the fusion of all the modalities available at a particular instant [8] [10] [9]. This has prompted researchers in the computer vision community to make use of the rich multimedia information (especially audio) in a video sequence for video surveillance.

A greater understanding of the human perception has led researchers to use cross modality in numerous projects for increase in accuracy and reliability. Lo and Goubran proposed a method for performing audio-video talker localization [3] that explores the reliability of the individual localization estimates such as audio, motion detection, and skin color detection. Lately, surveillance systems are using both audio and video sensors to reveal and track the presence of an intruder. The system described in [5] is composed of

a mobile agent and several static agents cooperating in the tracking task. In [16] a content based video parsing and indexing method is presented that analyzes both information sources (audio and video) based on their inter-relations and synergy to extract high-level semantic information. Speaker localisation using audio-video cues at signal level has been explored in [14] and [11]. Both use mutual information between signals in audio and video domains. Kidron, Schechner and Elad have associated sound to the relevant pixels in the video [4] using canonical correlation analysis. In this paper we look in the problem of separating more than one concurrent audio and video events using a feature based approach based on single audio and a single video sensor. The use of higher level primitives and grouping makes audio-video association more robust. The approach also works well in a cluttered environment where more than one object exists and associates sound to the corresponding object at a particular instant.

2. Perceptual Organization of Sound and Video

Humans use their sense of hearing to understand the properties of sound-producing events. In a natural listening environment the acoustic energies produced by several events are mixed, at the listener's ears, with energy arising from other concurrent events. "Auditory scene analysis" (ASA) is a process in which the auditory system takes the mixture of sound that it derives from a complex natural environment and sorts it into packages of acoustic evidence in which each package probably has arisen from a single source of sound [1]. ASA provides answers to how the brain can build separate perceptual descriptions of sound-generating events despite the mixing of evidence. The first thing it does is to analyze the incoming array of sound into a large number of frequency components. Then, by putting together the right set of frequency components over time, a signal is recognized. Bergman [1] showed the similarities between the Gestalt principles in vision and audition. Just as grouping by proximity functions operates in visual space it also operates auditory pitch. McPherson, Ciocca and Bergman [12] have shown that good continuation operates in audition in an analogous way to vision. The concept of amodal completion as it is used in vision [6] has been

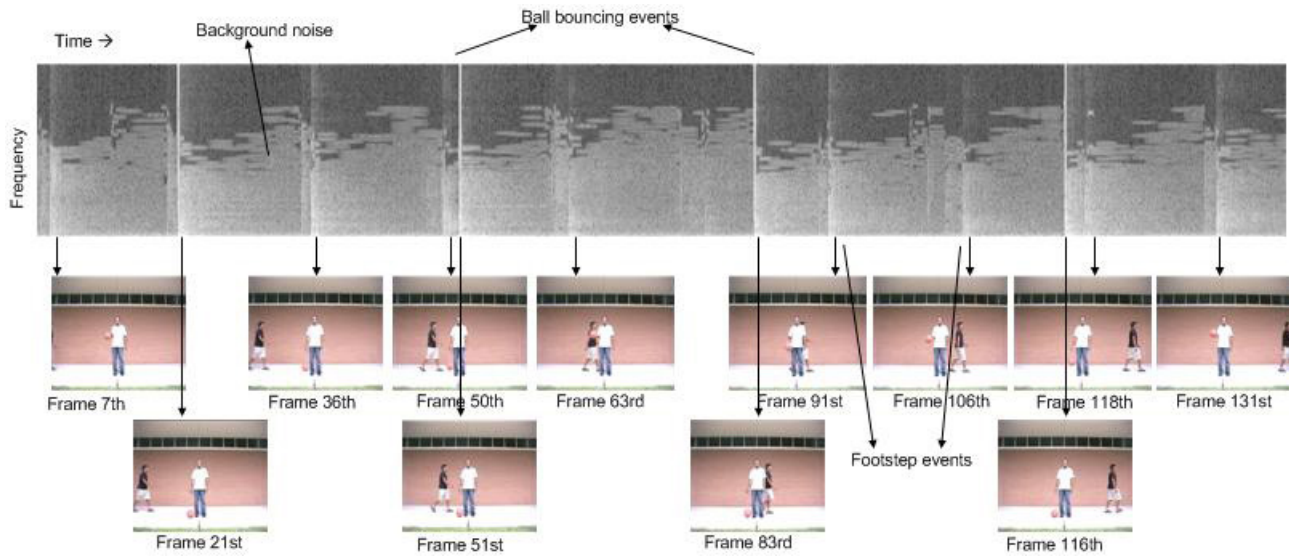


Figure 1. (Top) Spectrogram of the sound waveform received. (Bottom) Corresponding frames for each event in the spectrogram

given a number of different names in audition: the acoustic tunnel [7] effect, perceptual restoration [15] and the continuity effect [1]. Since all these phenomenon abide by the same laws of grouping and organization, a framework which accounts for these needs to be used.

A perceptual object is that which is susceptible to figure-ground segregation. Early processing produces elements that require grouping. Grouping occurs following the principles described by the Gestalt psychologists; it produces Gestalts or perceptual organizations, which are also putative perceptual objects. Attention selects one putative object and relegates all other information to ground [13]. There is little doubt that grouping and figure-ground segregation describe processes that are meaningful for auditory perception. Another phenomenon that characterizes visual objects is the formation and assignment of edges. However, edges in audio seems to be an abstract concept. Kubovy and Valkenburg [13] developed the theory of indispensable attributes (TIA) which states that, in vision, objects are formed in space-time domain, however auditory objects are formed in pitch-time domain. Imagine presenting to an observer two spots of light on a surface. Both of them yellow and they coincide; the observer will report one light. Now suppose we change the color of the lights, so that one spot is blue and other is yellow, but they still coincide; the observer will report one white light. For the observer to see more than one light, they must occupy different spatial location. Now, imagine simultaneously playing two 440Hz sounds for a listener. Both of them played over the same loud speaker; the listener will report hearing one sound. Now suppose we

play these two sounds over two loudspeakers; the listener will still report hearing one sound. For the listener to report more than one sound, they must be separated in frequency. Thus, pitch separation is an indispensable attribute for audio perception. By analogous argument time is an indispensable attribute for both vision and audition. The TIA thus forms a heuristic tool for extending theories of visual perception into the domain of auditory perception.

3. Grouping Sound Events

As explained in the previous section, to process audio using ASA techniques we need a framework in the pitch-time domain. We use the spectrogram to represent audio. The spectrogram uses a slightly different form of DFT called the short time fourier transform (STFT). The STFT is a formulation that can represent sequences of any length by breaking them into shorter blocks, or frames, and applying the DFT to each block. Digitally sampled data, in the time domain, is broken up into frames, which usually overlap, and Fourier transformed to calculate the magnitude of the frequency spectrum for each frame. Each frame then corresponds to a vertical line in the image; a measurement of magnitude versus frequency for a specific moment in time. The spectrums or time plots are then "laid side by side" to form the image. The horizontal axis represents time, the vertical axis is frequency, and the intensity of each point in the image represents amplitude of a particular frequency at a particular time.

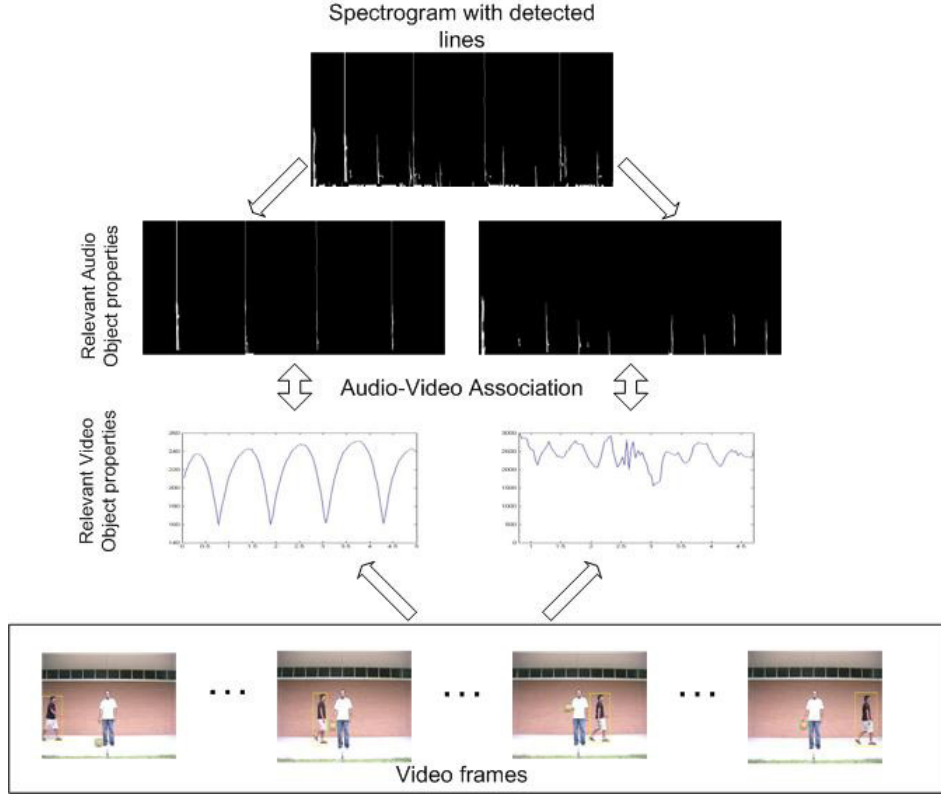


Figure 2. Top row shows the lines detected from the spectrogram of the sound wave from which we obtain audio objects shown in the second row. Third row shows the periodicity curves of objects in the video extracted from video frames shown in the last row.

A common observation from inspection of several spectrograms from outdoor noises is that significant audio events such as a walking person, bouncing ball, car horn etc. produce straight lines in the spectrogram image (Top of Fig 1). We can detect these lines and group those that belong to a specific event. In this work we use a line detection algorithm based on [2] to extract these straight lines (First row of Fig 2). Once we obtain the spectrogram image with the lines representing only the significant audio events, the orientation and centroid of each line is calculated. The orientation is the angle between the x -axis and the major axis of the ellipse that has the same second moments as the line. The centroid is the center of mass of the region. The lines are then grouped into four groups based on orientation values; $\theta_1(0^\circ-30^\circ$ and $150^\circ-180^\circ)$, $\theta_2(30^\circ-60^\circ)$, $\theta_3(60^\circ-120^\circ)$, $\theta_4(120^\circ-150^\circ)$.

Though this step will group all lines with similar orientation, lines with different pitch information may also be grouped together. To avoid this we further group lines in each orientation group based on the centroid. The coordinates of the centroids of lines are clustered using a sim-

ple k -means algorithm. This location based grouping extracts lines that are close together and along the same coordinate axis signifying lines belonging to the same auditory event. This process however still results in groups that are formed due to low frequency noise in the spectrogram. These groups are ignored during the association stage where they do not find any video object to associate themselves to.

Once we isolate the auditory events from the spectrogram it is easy to estimate its periodicity. The peaks in the auditory signal can be obtained by finding the local maxims in the column summation array of the spectrogram. The time difference between each peak is obtained by subtracting consecutive peak times. The average of these differences gives us the periodicity of the signal each group of elementary audio events.

4. Grouping of Video Events

Moving objects can be detected by background subtraction. Once we subtract background frame from each frame we threshold the image to get a binary version with just the

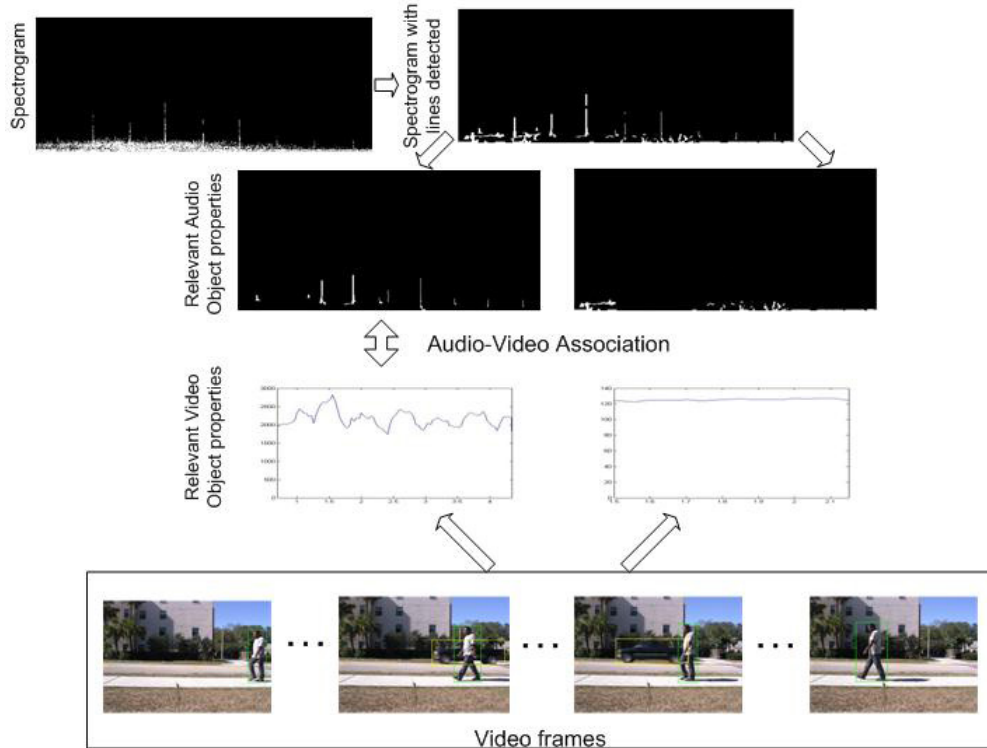


Figure 3. Top row shows the spectrogram on the left side and lines detected on the right side. Two of the audio objects are shown in the third row. Fourth row shows the periodicity curves of objects in the video extracted from video frames shown in the last row.

objects in motion. Video objects exhibit two types of periodicities: shape change periodicity or track periodicity. Periodic behaviour of objects such as humans exhibit shape change periodicity and can be estimated by measuring the change in the number of pixels in the bounding box in each frame. This is based on the fact that in a 2D image the number of pixels corresponding to the human decrease as both the legs come together and increase as the step is completed. The periodic behaviour of objects such as a bouncing ball can be calculated by tracking the centroid of the bounding box in each frame. Once we have the periodic curves we can find the periodicity by applying periodic transforms [17]. Periodicity Transforms decompose a data sequence into a sum of simple periodic sequences by projecting onto a set of periodic subspaces, leaving residuals whose periodicities have been removed. This decomposition is accomplished directly in terms of periodic sequences and not in terms of frequency or scale, as do the Fourier and Wavelet Transforms. Unlike most transforms, the set of basis vectors is not specified a priori, rather, the Periodicity Transform finds its own best set of basis elements. Though we estimate both the periodicities for each object, only one type of periodicity

will be relevant to an object. For example, the human might exhibit periodic motion by monitoring the change in pixels in a bounding box, but tracking the centroid will reveal only a straight line. The periodic transform will estimate periodicity as zero and hence we can ignore it.

5. Association

Audio and video are highly correlated, for example the sound of a ball bounce will be accompanied by an event in the video where the ball comes in contact with the floor (Frames 21,52,83,116 in Fig 1). Similarly, the sound of a footstep will be accompanied by the foot of a person hitting the floor (Frames 36,50,63,91,106,118,131 in Fig 1). We simply need to keep track of when we hear the sound and in video we need keep track of when the ball or the foot comes in contact with the floor. Essentially, the periodicities of objects both in audio and video should be similar. Video will have exactly the same number of periodicities as the number of objects in the scene. However, in audio due to noise, our procedure might pick up false periodicities. To eliminate these we simply do a percentage difference check between

the video and audio periodicities. The lowest differences will give us the video object and audio object which are most similar.

6. Results

First, we consider a scenario of a bouncing ball and person walking (Fig 1). The detected lines from the spectrogram are shown in the first row of figure 2. We obtain three auditory objects with periodicities 1.06, 0.8 and 0.46 seconds. Some of the relevant audio objects are shown in the second row. The first figure in the third row represents the periodicity curve of the bouncing ball. The periodicity curve of the human walking is shown in the second figure in the same row. The periodicities are 1.04 seconds and 0.52 seconds respectively. We calculate the difference percentages from these. Since we have two video objects, the smallest two percentages form the association. Hence, the bouncing ball video object one is associated with audio object three and human walking is associated with audio object one.

Now, we consider another scenario in which we have a walking person and car passing by in the background. The spectrogram of the audio is shown in the right side of the first row of figure 3. The lines detected from the spectrogram image is shown on the right side. Two of the audio objects are shown in the third row. The first audio object corresponds to the footsteps of the walking person. The second object is formed due to low frequency noise. However, no audio object corresponding to the sound of the car is formed. The video object of the walking person exhibits shape change periodicity but the motion of the car is uniform. During the association stage no audio object is associated to the moving car. We indicate this in figure 3 by not showing any bi-directional arrow for the second video object.

Spectrogram images of the isolated auditory events can also be used to regenerate the corresponding audio signal. By retaining only the columns in which at least one pixel is equal to one, in the original spectrogram matrix (obtained by applying fourier transform to the original wave form) and making all others zero, we can then apply a frame by frame inverse fourier transform and append the signal to obtain the corresponding sound.

7. Conclusion

In this paper, we present a technique by which we can associate sound to motion in the video. Specifically, we associate sound to objects exhibiting shape change or track periodicities. Our method uses a feature based approach which groups high level primitives, thus avoiding the noise associated with using low level features. This approach works

well even in a cluttered environment and can associate more than one sound to the respective objects at the same instant.

References

- [1] A S Bergman. *Auditory scene analysis: The perceptual organization of sound*. MIT Press, 1990.
- [2] C Stegner. An Unbiased detector of Curvilinear structures. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 20, pages 113–135, 1998.
- [3] D Lo and R A Goubran. Robust joint audio-video localization in video conferencing using reliability information. In *IEEE Transactions on instruments and measurements*, volume 20, pages 1132–1139, 2004.
- [4] E Kidron, Y Y Schechner and M Elad. Pixels that Sound. In *International conference on pattern recognition*, volume 1, pages 88–95, 2005.
- [5] E Menegatti, E Mumlo, M Nolich and E Pagello. A surveillance system based on audio and video sensory agents cooperating with a mobile robot. In *International Conference on Intelligent Autonomous Systems*, pages 335–343, 2004.
- [6] G Kanizsa. *Organization in vision: essays on Gestalt principle*. New York: Praeger, 1979.
- [7] G Vicario. The acoustic tunnel effect. *Rivista da Psicologia*, pages 41–52, 54, 1960.
- [8] H McGurk and J MacDonald. Hearing lips and seeing voices. *Nature*, pages 746–748, 1976.
- [9] J Vroomen and B de Gelder. Sound enhances visual perception: Cross-Modal effects of auditory organization on vision. *Journal of Experimental Psychology: Human perception and performance*, pages 1583–1590, 2000.
- [10] J Vroomen, P Bertelson and B de Gelder. A visual influence in the discrimination of auditory location. In *Proceedings of the International Conference on Auditory-Visual Speech processing*, pages 131–135, 1998.
- [11] J W Fisher and T Darrell. Speaker Association With Signal-Level Audiovisual Fusion. In *IEEE Transactions on multimedia*, volume 6, pages 406–413, 2004.
- [12] L McPherson, V Ciocca and A Bergman. Organization in audition by similarity in rate of change: evidence from tracking individual frequency glides in mixtures. *Perception and Psychophysics*, pages 269–278, 1994.
- [13] M Kubovy and D V Valkenburg. Auditory and visual objects. *Cognition*, 80:97–126, 2001.
- [14] R Cutler and L Davis. Look who’s talking: Speaker detection using video and audio correlation. In *IEEE International Conference on Multimedia and Expo*, volume 3, pages 1589–1592, 2000.
- [15] R M Warren. *Auditory perception: a new synthesis*. New York: Pergamon Press, 1982.
- [16] S Tsekeridou and I Pitas. Content-Based video parsing and indexing based on audio-visual interaction. In *IEEE Transactions on circuits and systems for video technology*, volume 11, pages 522–535, 2001.
- [17] W A Sethares and T W Staley. Periodicity Transforms. In *IEEE Transactions on Signal Processing*, volume 47, pages 2953–2964, 1999.