Automated Computational Matching of Video and Audio Content

Daniel S. Soper [(1)]

[(1)] *Department of Information Systems & Decision Sciences, California State University Fullerton, CA 92831 USA*
+1.657.278.7270 *dsoper@fullerton.edu*

1. Introduction – This research project inquires into whether it is currently possible for a computer program to automatically select music that matches the content of a video clip. To explore this possibility, it was necessary to integrate and expand upon techniques originating from several different fields, including computer vision, audio processing, and pattern analysis. Using state-of-the-art object tracking algorithms in conjunction with audio waveform analysis and error minimization-based pattern matching, the results indicate that computers can indeed effectively select music to accompany short video clips.

2. Methods - Five public-domain video clips dating from the 1930s to the 1950s were chosen as video input for the experiment. The videos all depicted scenes of people dancing or playing instruments since such videos represent natural targets for musical accompaniment. 100 modern, public-domain "electronica" songs were chosen as audio input for the experiment. The input songs were characterized by a wide variety of tempos, instruments, and other musical properties. A modified, state-of-the-art Discriminative Scale Space Tracker (DSST) was implemented (see Figure 1) in order to track and analyse the movements of people in the videos [1], and create a motion profile for each video. Next, the mean motion and its associated standard deviation were used to compute a motion coefficient of variation (mCOV) for each video. Each video's original audio was then run through a beats-per-minute (BPM) analyser in order to establish the tempo of the audio [2]. Using linear modelling, the mCOV was found to be an excellent predictor of BPM. The resulting linear equation was then used to locate songs in the input set whose tempos matched each video. After identifying the input songs whose tempos matched the motion in a video, the final step was to isolate the segment of audio within those songs that best matched the video's motion profile. For this purpose, each input song was downsampled such that its sampling rate matched the video framerate. The video's motion profile waveform was then iteratively compared to every possible audio segment waveform of the same duration in order to identify the audio segment which most closely matched the video's motion profile (see Figure 2).

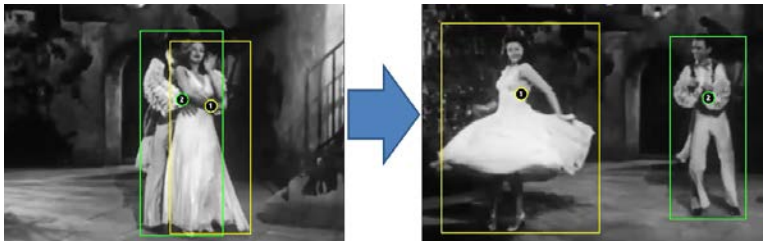

Figure 1. Automated tracking of subject motion using a Discriminative Scale Space Tracker.

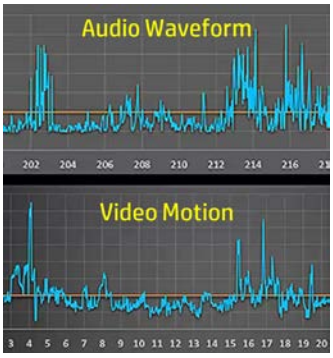

Figure 2. Matching audio waveforms to video motion.

3. Results and Conclusion - The resulting video clips with their computationally selected audio accompaniment were played for a group of 36 graduate students. Depending on the specific video clip, between 94.4% and 100% of these subjects indicated that the selected audio was an excellent match for the video. It is thus possible to conclude that computers can be programmed to effectively select audio which matches the motion in a video clip. The results from the experiment will be demonstrated at the conference.

4. References

[1] Danelljan, M., Häger, G., Khan, F., & Felsberg, M. *Accurate scale estimation for robust visual tracking*. British Machine Vision Conference, Nottingham, September 1-5, 2014.
[2] Seyerlehner, K., Widmer, G., & Schnitzer, D. *From Rhythm Patterns to Perceived Tempo*. International Society for Music Information Retrieval, (2007) p.519.