1. Motivation

Goal: to find derivative works of popular songs posted online.

Music videos and their derivatives are hugely popular on YouTube and other services.

Audio Content (AC) and Video Content (VC) are independent, and both important:

- **AC**: Original audio
- **VC**: Official music video
- **AC**: Cover (new arrangement)
- **VC**: Live (performance)
- **AC**: Remix (samples orig.)
- **VC**: Still image
- **AC**: Original audio
- **VC**: Dance performance

Problem: text search for derivatives gives many errors: a search for “covers” often gives remixes; a search for “live” can give covers; etc.

Solution: build a system to re-classify search results based on search rank, text [i.e., video title], audio and video features.

2. Related work

Techniques exist to identify specific types of derivatives:

- **Fingerprinting**: peaks in audio signal spectrum are robust to noise and can be used to identify exact copies of audio.
- **Remix detection**: look for any matching fingerprints, possibly time- or frequency-distorted.
- **Cover song detection**: characterize chord and/or melody structure, then look for matches.
- **Live song identification**: use a variation on fingerprinting that is feasible if artist is known.

But these methods were not designed and are not guaranteed to distinguish between types of derivatives.

3. Proposed system

**Inputs:** keyword list and query song/artist

**Keywords**
- “choreo”
- “concert”
- “dance”
- “guitar”
- “lyrics”
- “remix”
- “tutorial”

**Query**

- “Gotye Featuring Kimbra, ‘Somebody That I Used To Know’”
- “AC: miles davis, ‘take five’”
- “VC: dance/cover”

**Baseline**

- SVM with a multi-modal set of four features:
  - AC accuracy
  - VC accuracy

**STV**

- Initial predictions
  - SVM classifier
  - “STV” classifier

- Extract features:
  - Search features
  - Video features
  - “STV” features

- Phase 1: make initial prediction of AC and VC for each video.

- Treat the highest-confidence examples for each AC category as class exemplars.

- Refined predictions
  - SVM classifier
  - “STV” classifier

- Extract audio features (compare audio to class exemplars)

- Phase 2: compare audio files to class exemplars, and obtain refined predictions.

4. Features

To estimate the AC and VC of videos, we use an SVM with a multi-modal set of four features:

**Search**: a video can appear in several keyword searches. We use a video’s rank in each search. E.g., a dance video might have ranks:

- Keyword: main choreo concert cover dance guitar ..
- Rank: 22 1 361 3 ...

**Text**: words that appear in many titles are likely to relate to AC or VC: e.g., “acoustic”→“cover”, “tour”→“live”, “vs”→“remix”, etc. From an unlabelled set of ~150,000 video titles, we learn latent topics. We can then convert each title to a “topic strength” vector. To help the learning process, we detect terms from dictionaries of places, names, instruments and genres.

**Video**: common video features (e.g., brightness, colour variance, optical flow). To detect lyrics, we do text-recognition using Tesseract.

**Audio**: we use audio fingerprinting to match 10-second snippets to the original song. The distribution and quality of matches found is different for different types of derivatives.

5. Evaluation

**Data**: we discovered 160,000 derivatives of the Billboard Top 100 songs of 2012. We labeled 562 videos related to 10 unique songs, and used the rest (titles & search ranks) for unsupervised training.

**Ground truth**: manually annotated AC and VC:

- 6 AC categories: official, cover, instrumental, live, remix, tutorial.
- 9 VC categories: official, dance, karaoke, live, lyrics, slideshow, still image, tutorial, other.

**Baseline**: a decision tree (DT) using peak search rank from YouTube searches. E.g., if video has the highest rank for the “remix” keyword search, then AC→remix and VC→still.

**Results**:

- Text features alone match the performance of the YouTube baseline.
- Combining all features (STV) in two-phase approach, we surpass baseline by 10%.
- Our system is also more robust than baseline: classification of deep search hits still accurate.