QUERY BY VIDEO: CROSS-MODAL MUSIC RETRIEVAL

Bochen Li
University of Rochester
Rochester, NY, USA

Aparna Kumar
Spotify
New York, NY, USA

ABSTRACT

Cross-modal retrieval learns the relationship between the two types of data in a common space so that an input from one modality can retrieve data from a different modality. We focus on modeling the relationship between two highly diverse data, music and real-world videos. We learn cross-modal embeddings using a two-stream network trained with music-video pairs. Each branch takes one modality as the input and it is constrained with emotion tags. Then the constraints allow the cross-modal embeddings to be learned with significantly fewer music-video pairs. To retrieve music for an input video, the trained model ranks tracks in the music database by cross-modal distances to the query video. Quantitative evaluations show high accuracy of audio/video emotion tagging when evaluated on each branch independently and high performance for cross-modal music retrieval. We also present cross-modal music retrieval experiments on Spotify music using user-generated videos from Instagram and Youtube as queries, and subjective evaluations show that the proposed model can retrieve relevant music. We present the music retrieval results at: `http://www.ece.rochester.edu/~bli23/projects/query.html`.

1. INTRODUCTION

Music retrieval has been explored for many cross-domain inputs such as text [27], image [5], location [41], video [32], vocal imitation [42], and sheet music [29]. To our knowledge there are few reports focusing on cross-modal music retrieval given videos from unconstrained sources. With the proliferation of smart phones, people capture short videos to communicate moments from their everyday lives. Learning relationships between music and real-world videos has many applications including novel music query scenarios where a playlist is recommended to fit user's surrounding scenes, or automatically soundtrack selection to complement and enhance visual messages on social media, e.g., Snapchat, Instagram, and Facebook.

Real-world videos can contain any form of video including edited or raw content, and music is an inherently diverse content as well [8]. Thus associating music with such videos is more challenging than common sounds and objects, e.g., barking to a dog [1], which has explicit connection on semantic level. One way to bridge real-world videos and music is via elicited emotions. Previous work addresses this problem after recognizing each modality independently as hand-labeled emotion features [5, 32], but this is not sufficient since hand-labeled emotions (e.g., several human-defined emotion tags) are prone to bias and subjectivity [39], and bottleneck each modality into a limited non-learnable space. Thus no scalable solution has been proposed to query from large music databases. Later a two-stream network structure is proposed for music query by music videos [12], where cross-modal embeddings are learned directly from music-video pairs. However, it requires intensive training on music videos (MV) where the videos were originally created for specific songs, and the music retrieval performances given videos from more varieties of sources are not systematically evaluated.

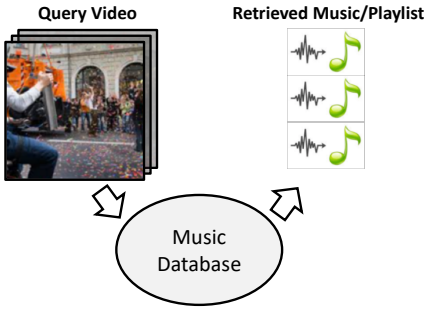

Figure 1. A large music database can be queried by real-world videos from unconstrained sources.

In this paper, we address the music retrieval task on videos in the wild (Figure 1). Different from previous work that models each modality independently, we propose a two-stream network structure to learn the cross-modal distance in an end-to-end fashion using music-video pairs, while emotion tags are applied on each branch to form latent emotion space. Each branch is pre-trained as audio/video emotion tagging sub-network before feeding music-video pairs to both for cross-modal distance learning. This strategy requires fewer music-video pairs for training, and makes it possible to collect crowdsourced pairs of music and videos from independent sources. Note that the tags are only used during training phase to facilitate the convergence of cross-modal distance learning, and not necessary during model inference of music query. When a video queries the system, the model ranks every item of

an existing music database by Euclidean distance to the input video on the cross-modal embedding. The top ranked results represent the best matches to the input.

The main contributions of this paper are :

- A first system to address music retrieval from videos in the wild via learnable emotion space.
- A two-stream network structure and training strategy with emotion tags as joint constraints to learn cross-modal embeddings from fewer music-video pairs.
- Subjective evaluations showing promising retrieval results on real-world datasets.

2. RELATED WORK

2.1 Music, Videos and Emotions

The emotions associated with music and videos have been thoroughly studied. It has been suggested that emotions are one of the primary reasons people engage with music [15], and psychological studies reveal that people have emotional reactions on visual stimuli as well [7]. Therefore a natural way to retrieve music for videos is through the associations with emotion.

Categorical and dimensional representations have been used to represent emotion in music [18]. Discrete categorical tags include terms such as *calmness*, *sadness*, *anger*, and more. Gracenote [1] has performed a major effort around tagging the mood in music and provides mood taxonomies consisting of over 300 categories organized hierarchically. One work finds that the number of mood categories does not reflect the richness of emotions perceived by humans, or the taxonomy is inherently ambiguous [15]. Dimensional labels typically represent music on a 2-D plane of valence and arousal [30]. This continuous representation does not have the taxonomy problem, but has trouble distinguishing some psychology and emotion concepts such as *nostalgia*.

Emotion associated with images and videos have been also represented categorically [44] and dimensionally [24], similar to music. Seven emotion tags have been associated with videos of facial expressions [16, 17]. Eight basic emotion tags, with 3 variations on each tag are introduced for labeling unconstrained videos [38]. Movie scenes have been characterized in the valence-arousal space [3]. In [11], "Dominance" is introduced as an additional dimension for characterizing video emotions.

2.2 Cross-modal Audio-Visual Retrieval

Cross-modal retrieval has received increasing attention in the recent years. One work proposes a two-stream network structure for audio-vision cross-modal retrieval of common objects and their respective sounds, such as an image of a clock paired with the sound of an alarm [1]. This work has curated a large training dataset of common objects and sounds from publicly available sources. The cross-modal correspondence is learned from audio-visual pairs. Similar work has been described with additional modalities of text [2] and speech [26].

Related work for music includes cross-modal localization [43], association [20, 22], and generation [21] of music performances. Earlier work for cross-modal music retrieval involves extracting distance measurement between low-level features from video and music segments [40]. Some approaches synchronize video and music after representing each modality as sequence of 2-D valence-arousal features [23, 31, 32]. One work uses stochastic emotion space to bridge video and music [36], and another reports recommended music for still photo albums by defining a cross-modal graph on which synsets of mood tags from images and music are associated [5]. Pairing in these approaches recognizes each modality as explicit symbolic representations independently (e.g., hand-labeled emotions) before learning the association. Also, these systems mostly emphasize temporal inter-dependence, focusing on pairing a soundtrack to match the visual event with less demand on learning deep semantic representations on emotions.

Learning cross-modal embeddings end-to-end using cross-modal pairs could result in a deep representations of the relationships and improved performance at scale in a music retrieval setting. As presented in [12], music/video cross-modal retrieval has been modeled by learning from music-video pairs and presented on music and their respective music videos, where the videos were intentionally created as MV. Without constraining the cross-modal learning space, it requires intensive training on music-video pairs, e.g., existing music videos, and is not systematically evaluated on the retrieval results for videos from more sources. In this paper we also learn the embedding space in an end-to-end fashion using pairs of video and music, but constrain the learning space with emotion tags to form latent emotion space for each modality.

3. APPROACH

3.1 Network Architecture

3.1.1 Video Branch

The video branch consists of a feature extraction module followed by fully-connected layers for emotion tagging, left stream in Figure 2. We use pre-trained Inflated-3D model (I3D) [4] as the visual feature extractor. I3D was originally proposed for human action recognition from videos and was trained on the Kinetics dataset [4]. This pre-trained network has been successfully used for other video understanding tasks such as video captioning and audio-visual localization [34].

We input only the RGB frames and ignore optical flow. The system outputs a concatenation of the inception modules. We take a global average pool resulting in a 1024-D feature vector for each whole video of any duration. Next, we add fully connected layers. Each layer is followed by a ReLU nonlinearity, except that the output layer is followed by a sigmoid nonlinearity. Input video frames are resized to 224×224, and the RGB values are normalized to [-1,

[1] www.gracenote.com

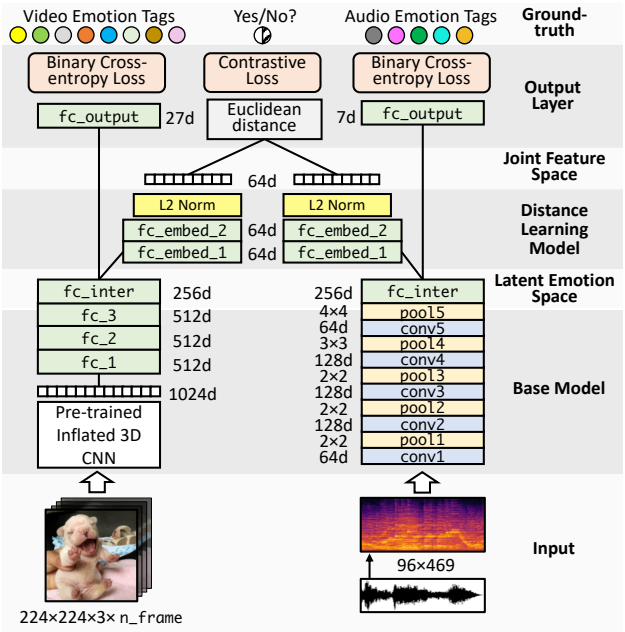

Figure 2. The two-stream network architecture with emotion constraints on each branch. The dimensions of fully-connected layer (fc), channel numbers of convolutional layer (conv), and pool sizes are marked aside each block. The kernel size for all the convolutional layers is 3×3.

1]. The video network is pre-trained with 27 emotion tags using binary cross-entropy loss.

3.1.2 Audio Branch

The audio branch consists of a ConvNet structure as a general audio tagging framework analogue to [6], right stream in Figure 2. Each convolutional layer is followed by a batch normalization and ReLU nonlinearity, and the output fully-connected layer has a sigmoid nonlinearity. Audio is trimmed to 10-sec with a sample rate of 12 kHz. Log mel-spectrograms are computed from input audio with a frame length of 42.7 ms (with 50% overlapping) and 96 mel-scale filter banks. The audio network is pre-trained with 7 emotion tags using binary cross-entropy loss.

3.1.3 Cross-modal distance learning

The cross-modal distance learning network is designed to embed the video and audio into the cross-modal embeddings (i.e., the joint feature space in Figure 2) so they can be directly compared as vector distance. This network takes the two 256-D penultimate layers from the video and audio branches to predict if the input music-video pair match, as a binary classification problem. The 256-D layers represent latent emotion space that is learned from the training pairs. The classifier is trained with the contrastive loss [10] on the Euclidean distance between each modality's 64-D cross-modal embedding, after L2 normalization.

3.2 Training

We first pre-train the audio and video branches independently as multi-label classifiers to predict emotion tags for each modality. The training stops when validation loss does not decrease for 5 consecutive epochs. Then the cross-modal distance learning framework is trained jointly while each branch predicts emotion tags. The network is jointly constrained and the three loss functions are equally weighted. This strategy constrains the learning space using emotion tags, and enables cross-modal distance learning from fewer music-video pairs. We use Adam optimizer [19], a stochastic gradient descent method, to minimize all the loss functions. When the model is trained, in practice all tracks in a music catalog are indexed by the embedding vector from the cross-modal joint feature space. Given any query video, the tracks in the database can be ranked by the Euclidean distance to the embedding vector calculated from the video. This creates a fast retrieval setup for large catalogs.

4. DATA

4.1 Training Data

The audio and video branches are pre-trained on independent music and video datasets with emotion labels. To train the cross-modal network we reuse the data from different modalities to create music-video pairs according to crowdsourced annotations about how well each pair matches.

4.1.1 AudioSet

We use the AudioSet [9] to pre-train the audio branch. AudioSet has human-labeled 10-second sound clips drawn from YouTube videos. We use data from "Music Mood" Ontology which contains music excerpts that are labeled with one of 7 music mood categories. We use AudioSet's official *Unbalanced Train* data where we randomly sample roughly 800 clips from each category for a total of 5.6K samples. We split it randomly where 80% is used for training and 20% is used for validation. We use AudioSet's official *Eval* data as the test set which consists of 354 samples, roughly 60 for each class, barring invalid download links. We do not use the videos from AudioSet for cross-modal distance learning because most contain a specific type of edited content, which is not suitable for the objective of retrieving music for videos from unconstrained sources.

4.1.2 Cowen2017

We use the Cowen2017 dataset [7] to pre-train the video branch. The dataset includes over 2K data samples including video clips from daily life, movies, cartoons, game scenes, artistic work, and more. Each video is annotated by several subjects who could select up to 27 emotion tags for each video. The annotations are aggregated so that each video's label is 27 emotion tags with a confidence value between 0 and 1. We split it randomly where and 80% is used for training and 20% is used to create the test set.

4.1.3 Music-video Pairs

To our knowledge there are no publicly available datasets that connect diverse videos with music, and contain the

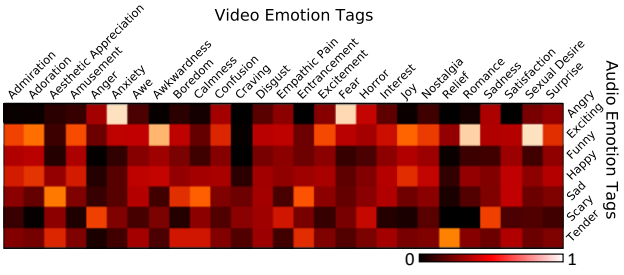

Figure 3. Visualizations of the emotion tags for the annotated music-video pairs from crowdsourced annotations. The color shows normalized counts.

respective emotion labels. Our goal is to construct a generalizable dataset that matches samples from one modality to the other while all samples have emotion labels. So we apply crowdsourcing to create music-video pairs from the same samples in AudioSet and Cowen2017. We follow the respective splits of training and test set as before so clips from one set do not appear in the other.

Training a cross-modal network requires positive and negative music-video pairs. The positive pairs are created by collecting binary crowdsourced judgments for randomly paired music and video clips until we have 1000 matches. Detailed crowdsourcing setup is described in Section 4.3. The negative samples are created from random un-annotated music-video pairs. In total we have class-balanced training set of 2000 pairs. We then perform the same process to collect another 1000 pairs on the test set. In Figure 3, we present a heatmap visualization of the relationship between the emotion tags of the positive audio-video pairs from the two modalities.

4.2 Real-world Data

To estimate how the system will perform on real-world data we curate videos and music from popular social media and music streaming platforms.

4.2.1 Spotify's Popular Music

We create a dataset of popular music from Spotify, an international music streaming platform. We identify popular Gracenote level 1 worldwide genres where at least 1000 tracks are streamed per day on Spotify. From each of the 30 most popular genres we select 40 of the most popular songs. The audio is downloaded from Spotify, which results in 1195 music clips.

4.2.2 The Moments in Time Dataset

We use video clips from the *Moments in Time* dataset [28] where each clip is a 3-second video snippet. The dataset was created for the tasks of action recognition and event understanding. We pick the first 100 moment categories (sorted alphabetically) from the *Moments in Time Mini* (a subset). From each category we select the first 5 video samples, totaling 500 video clips as the query videos.

4.2.3 Instagram Videos

Instagram is a social media platform for sharing photos and short videos. From a new account without search history we curate the top 20 videos from common photo post categories [13] : *Friends*, *Food*, *Gadget*, *Pets*, *Activities*, *Selfies*, *Fashion*, and we exclude *Captioned Photos* because the text may bias annotators' judgments and the system is currently not trained to process text. This results in 140 user uploaded short video clips.

4.3 Crowdsourcing Setup

Crowdsourced judgments are collected to create music-video pairs in the training and test datasets for cross-modal distance learning, and for subjective evaluations of music retrieval performance on real-world datasets. Experiments are run on the Figure-eight [2] platform which minimizes malicious activity during annotations and ensure high quality judgments for researchers.

Annotators are sourced from an international pool and each annotator is allowed to answer at most 10 questions, so that relevance judgments would not be overfit to any small group. Every question is randomly presented to at least 3 annotators. If the agreement among annotations is less than 65% per question, the number of annotators is dynamically increased up to 5 or until there is at least 65% agreement. Annotators are instructed to "listen in a quiet place, wear headphones, and watch the entire clip". The instructions for each audio-video pair are: "Please tell us if there is a common emotion theme in the video and the music, try not to focus on whether you like the music or the video." Possible responses are: "yes they match", "no they do not match", and "I am not sure".

To avoid biasing the pool of contributors we do not use gold standard screening questions that resemble the annotation questions, a common practice on Figure-eight. Instead to monitor annotation quality and attentiveness, we monitor whether any annotator's responses consistently deviate from the responses of other annotators. If an annotator's responses are different from the average annotation in more than 3 questions we flag that individual to analyze their contributions. We do not find any annotators fall into this group. In total we have collected thousands of annotations from subjects from 12 countries. Approximately 71% of the questions have greater than 66% agreement on "yes" and "no" responses. The remaining has responses split between "yes", "no", and "not sure", and for this work we consider the responses as "no" because annotators do not perceive relevance.

5. EXPERIMENTS

The model performance is evaluated numerically on the tasks of 1) predicting emotion tags from each branch independently, 2) predicting if the input audio-video pairs match, and 3) cross-modal music retrieval, on the annotated datasets. The music retrieval performances are also

[2] https://www.figure-eight.com

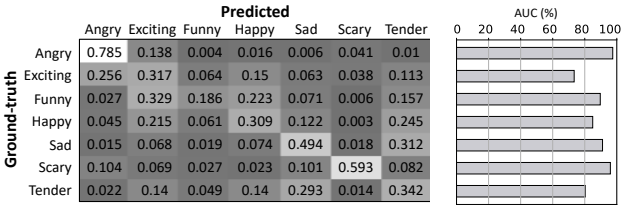

Figure 4. Performance of the audio branch for predicting emotion tags after pre-training. Results are presented as confusion matrix (left), and AUC on each category (right).

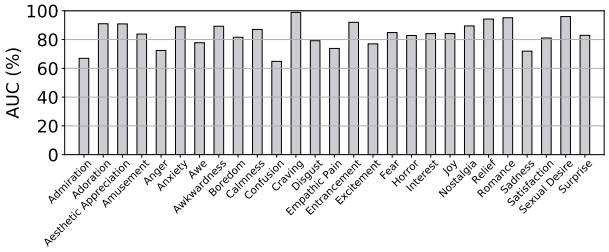

Figure 5. Performance of the video branch for predicting emotion tags after pre-training. AUCs are calculated using a score margin at 0.25.

evaluated on real-world data using crowdsourced subjective judgments.

5.1 Predicting Audio Emotion Tags

We evaluate the pre-trained audio branch using the labeled heldout data from AudioSet. The music emotion tagging result is evaluated using multi-class classification metrics. Figure 4 presents the confusion matrix, where *Angry* and *Scary* are likely to get high true positive rates, while the boundaries between *Exciting*, *Funny*, and *Happy* are blurred. We observe that angry music is generally noisy with strong percussion, while scary music has strong inharmonic components, as more distinct characteristics.

We also evaluate the model using the area under the receiver operating characteristic curve (AUC), a statistical metric that summarizes model's performance regardless of classification threshold. Equivalently, it measures the performance of binary classifiers (measuring each tag independently) by ranking scores, i.e., the probability that a randomly chosen positive is ranked ahead of a randomly chosen negative. The performance on each emotion tag is shown in Figure 4, with an average of **87.88**%.

5.2 Predicting Video Emotion Tags

The pre-trained video tagging branch is evaluated using heldout data from Cowen2017. Similar to scored AUCs [35, 37], we set a score margin on soft ground truth labels to report the performance. This metric assesses how well relative differences between video samples in the dataset can be predicted. It compares the sign of the differences between any two predictions to sign of the differences of the respective ground truth ones. Performance is measured only when the two data points have sufficiently large ground truth differences, e.g., 0.25 as used here. The average AUC for all tags is **83.79**%, as shown in figure 5 on each tag.

Method	2-stream	emotion [32]	emtoion [5]	proposed
Result	38.1%	36.0%	46.8%	**68.0**%

Table 1. Music retrieval performances on the labeled datasets compared with three baseline methods.

5.3 Predicting Audio-video Pairs

We also evaluate the cross-modal network to understand overall performance on cross-modal distance learning and the effects of the emotion tags which constrain the video and audio branches during cross-modal training. The cross-modal network predicts the input audio-video pairs as either positive or negative (matched or not) on the 1000 pairs from test set, and is evaluated as a binary classifier with a threshold of 0.5. We create a baseline model with the same two-stream network structure but without pre-training the branches on emotion tags or joint loss functions. The two models are trained and evaluated with the same dataset, described in Section 4.1.3. The accuracy of the proposed model is **79.00**% while the baseline achieves **63.30**%. Note that this baseline system is a general two-stream cross-modal distance learning network, e.g., [1], which usually requires intensive training on a large number of training pairs. The results indicate that pre-training and joint constraints on emotion tags is important for cross-modal distance learning when the training data is limited and the task includes data with highly diverse.

5.4 Cross-modal Music Retrieval

We reuse the heldout data from Cowen2017 and Audioset as query videos and the pool of music, respectively, where for each query video there are on-average 16.2 music samples from the pool are annotated as ground-truth retrieval. The music retrieval performance is evaluated by counting the number of videos that can retrieve a relevant song. For each query video only the top retrieval is considered, after ranking all the 354 music tracks. The proposed model retrieves relevant music for **68.0**% of query videos.

We also compare the performance to three baseline models, as presented in Table 1. The first baseline model is the same baseline as illustrated in Section 5.3, which share the same network structure but without emotion tags and pre-traning to form latent emotion space. It only achieves satisfactory retrieval for **38.1**% of videos.

The second baseline models each modality as continuous emotion representation on the valence-arousal (V-A) space, analogue to [32]. We implement this by adapting the proposed model structure, where the output layer from each branch is replaced with 2-D states to represent valence and arousal constrained by mean squared error (MSE) from the ground-truth values as a regression problem, without the two-stream structure for learning the cross-modal embeddings. The audio model achieves R^2

Video source	Moments in Time	Instagram
Result	58.2%	64.3%

Table 2. Music retrieval performance on Spotify music using query videos collected from new sources.

statistics of 53%/36% for arousal/valence when trained and evaluated on the *1000Song* dataset [33]. The video model achieves R^2 of 75.04%/60.28% for arousal/valence when trained and evaluated on the AudioSet videos using annotated labels. Both models achieve higher performance than the original work on emotion prediction [32]. We map the two modalities in the A-V space and the model achieves relevant retrievals for **36.0**% videos.

The third baseline matches modalities according to hand-labeled emotion tags, analogue to [5]. To implement this we modify the model structure to build the cross-modal distance learning structure from the predicted emotion tags from each branch, instead of the 256-D latent emotion space. This model achieves relevant retrievals for **46.8**% videos.

These experiments show that the proposed approach outperforms three baseline solutions. Two of the baselines join the modalities directly on predicted emotion states: arousal-valence values or explicit emotion tags. It suggests that our model learns deeper relationships between the modalities in the cross-modal space. Comparing to the other baseline, the results indicate that when the model is not constrained with emotion tags, only 2000 audio-video pairs as training set is too small for the network to learn the cross-modal embeddings to represent underlying the relationships between the cross-modal inputs.

5.5 Performance on Real-world Data

We assess how well our proposed model works on real-world data by collecting human judgments using the crowdsourcing setup in section 4.3. We use the 500 samples from the *Moments in Time* dataset and 140 user-generated videos from Instagram to retrieve music from a pool of music clips downloaded from Spotify. The model can successfully retrieve music for **58.2**% and **64.3**% videos, respectively. Music retrieval performance is better on Instagram than *Moments in Time*.

Note that Instagram videos are uploaded by users to visually share an experience or a mood that incites an emotion [14]. Instead, *Moments in Time* was created to capture different actions objectively without capturing sentiment [28]. This difference may explain why performance is higher on the Instagram videos.

5.6 Qualitative Analyses

We qualitatively analyze the latent emotion space learned from the two-stream model. We take the latent emotion space from the audio branch and create a t-SNE visualization [25], as plotted in Figure 6, to study how the matched videos localize in this 2-D space. Each dot represents a music sample from AudioSet, and we color the ones with the original label "Tender" in grey. Among these samples, we randomly select some from different regions and they are presented using different colors and with the thumbnails of the paired videos from model output. It indicates that samples close together with similar granular emotions are usually associated with similar videos. For example, samples in yellow represent music that sounds serene or soothing, and associated with outdoor nature scenes. This suggests that the proposed framework constrained with emotion tags enables the model to learn a interpretable emotion space and cross-modal correspondence including nuances that are no represented in the original tags.

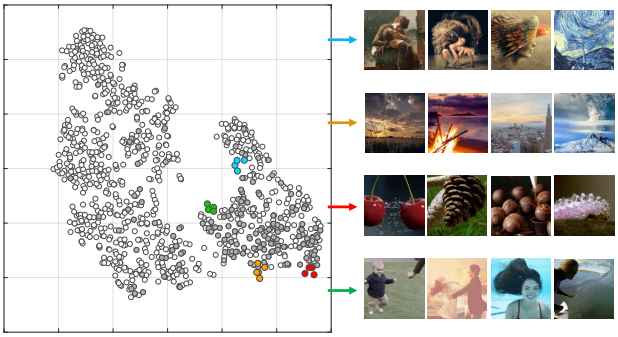

Figure 6. The t-SNE visualization of the latent emotion space from the audio branch. Samples from the "Tender" tag are in gray. Four randomly selected regions for tender music are presented in colors representing different emotion concepts: gloomy, ambient, delicate, sweet, each with the thumbnails of the paired videos displayed.

We also investigate some failure retrieval cases and find most are due to incorrect predictions of video emotions. Also, several retrievals are matched on emotions but mismatched on cultural signals. For example, a video with people bowing to Beyonce as she wears a crown is paired with music that sounds "mystical" or "heavenly", which is annotated as "mismatch" likely because annotators recognize Beyonce and they are expecting her music. If cultural signals are ignored these retrievals may have been reasonable matches. Overall, the results indicate that the system is effective for music retrieval, and we expect improvements by incorporating more signals such as culture or genres.

6. CONCLUSION

We have addressed the problem of music retrieval using real-world videos from unconstrained sources. We haved proved that emotion tags can constrain the learning space and enable cross-modal distance learning from fewer annotated cross-modal pairs. Experiments show that our model can retrieve promising results for user-generated query videos. As an application, this model can offer novel music query solutions for daily life videos which can enhance visual messages to make sharing more enjoyable. We also expect this work to have product implications in the music streaming business. In the future we plan to personalize the music that is retrieved for users' tastes.

7. REFERENCES

[1] Relja Arandjelović and Andrew Zisserman. Objects that sound. In *Proc. European Conference on Computer Vision (ECCV)*, 2018.

[2] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017.

[3] Yoann Baveye, Emmanuel Dellandrea, Christel Chamaret, and Liming Chen. Liris-accede: A video database for affective content analysis. *IEEE Transactions on Affective Computing*, 6(1):43–55, 2015.

[4] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[5] Jiansong Chao, Haofen Wang, Wenlei Zhou, Weinan Zhang, and Yong Yu. Tunesensor: A semantic-driven music recommendation service for digital photo albums. In *Proc. International Semantic Web Conference (ISWC)*, 2011.

[6] Keunwoo Choi, George Fazekas, and Mark Sandler. Automatic tagging using deep convolutional neural networks. In *Proc. International Society for Music Information Retrieval (ISMIR)*, 2016.

[7] Alan S Cowen and Dacher Keltner. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *National Academy of Sciences*, 114(38):E7900–E7909, 2017.

[8] J. Stephen Downie. Music information retrieval. *Annual Review of Information Science and Technology*, 37:295–340, 2003.

[9] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017.

[10] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proc. International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1735–1742, 2006.

[11] Alan Hanjalic and Li-Qun Xu. Affective video content representation and modeling. *IEEE Transactions Multimedia*, 7(1):143–154, 2005.

[12] Sungeun Hong, Woobin Im, and Hyun S Yang. Cbvmr: Content-based video-music retrieval using soft intra-modal structure constraint. In *Proc. ACM International Conference on Multimedia Retrieval*, pages 353–361, 2018.

[13] Yuheng Hu, Lydia Manikonda, Subbarao Kambhampati, et al. What we instagram: A first analysis of instagram photo content and user types. In *Proc. International Conference on Weblogs and Social Media (ICWSM)*, pages 595–598, 2014.

[14] Christina A Jackson and Andrew F Luchner. Self-presentation mediates the relationship between self-criticism and emotional response to instagram feedback. *Personality and Individual Differences*, 133:1–6, 2018.

[15] Patrik N Juslin and Petri Laukka. Expression, perception, and induction of musical emotions: A review and a questionnaire study of everyday listening. *Journal of New Music Research*, 33(3):217–238, 2004.

[16] Samira Ebrahimi Kahou, Xavier Bouthillier, Pascal Lamblin, Caglar Gulcehre, Vincent Michalski, Kishore Konda, Sébastien Jean, Pierre Froumenty, Yann Dauphin, Nicolas Boulanger-Lewandowski, et al. Emonets: Multimodal deep learning approaches for emotion recognition in video. *Journal on Multimodal User Interfaces*, 10(2):99–111, 2016.

[17] Heysem Kaya, Furkan Gürpınar, and Albert Ali Salah. Video-based emotion recognition in the wild using deep transfer learning and score fusion. *Image and Vision Computing*, 65:66–75, 2017.

[18] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *Proc. International Society for Music Information Retrieval (ISMIR)*, pages 255–266, 2010.

[19] D Kinga and J Ba Adam. A method for stochastic optimization. In *Proc. International Conference on Learning Representations (ICLR)*, volume 5, 2015.

[20] Bochen Li, Karthik Dinesh, Zhiyao Duan, and Gaurav Sharma. See and listen: Score-informed association of sound tracks to players in chamber music performance videos. In *Proc. International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2906–2910. IEEE, 2017.

[21] Bochen Li, Akira Maezawa, and Zhiyao Duan. Skeleton plays piano: Online generation of pianist body movements from midi performance. In *Proc. International Society for Music Information Retrieval (ISMIR)*, 2018.

[22] Bochen Li, Chenliang Xu, and Zhiyao Duan. Audio-visual source association for string ensembles through multi-modal vibrato analysis. In *Proc. Sound and Music Computing (SMC) Conference*, pages 159–166, 2017.

[23] Jen-Chun Lin, Wen-Li Wei, and Hsin-Min Wang. Emv-matchmaker: emotional temporal course modeling and

matching for automatic music video generation. In *Proc. ACM International Conference on Multimedia*, pages 899–902, 2015.

[24] Xin Lu, Poonam Suryanarayan, Reginald B Adams Jr, Jia Li, Michelle G Newman, and James Z Wang. On shape and the computability of emotions. In *Proc. ACM International Conference on Multimedia*, pages 229–238, 2012.

[25] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[26] Gou Mao, Yuan Yuan, and Lu Xiaoqiang. Deep cross-modal retrieval for remote sensing image and audio. In *Proc. IAPR Workshop on Pattern Recognition in Remote Sensing (PRRS)*, pages 1–7. IEEE, 2018.

[27] Brian McFee and Gert RG Lanckriet. The natural language of playlists. In *Proc. International Society for Music Information Retrieval (ISMIR)*, pages 537–541, 2011.

[28] Mathew Monfort, Bolei Zhou, Sarah Adel Bargal, Alex Andonian, Tom Yan, Kandan Ramakrishnan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *arXiv preprint arXiv:1801.03150*, 2018.

[29] Meinard Mueller, Andreas Arzt, Stefan Balke, Matthias Dorfer, and Gerhard Widmer. Cross-modal music retrieval and applications: An overview of key methodologies. *IEEE Signal Processing Magazine*, 36(1):52–62, 2019.

[30] James A Russell. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161, 1980.

[31] Shoto Sasaki, Tatsunori Hirai, Hayato Ohya, and Shigeo Morishima. Affective music recommendation system based on the mood of input video. In *International Conference on Multimedia Modeling*, pages 299–302. Springer, 2015.

[32] Ki-Ho Shin and In-Kwon Lee. Music synchronization with video using emotion similarity. In *Proc. International Conference on Big Data and Smart Computing (BigComp)*, pages 47–50, 2017.

[33] Mohammad Soleymani, Micheal N Caro, Erik M Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang. 1000 songs for emotional analysis of music. In *Proc. ACM international workshop on Crowdsourcing for multimedia*, pages 1–6. ACM, 2013.

[34] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proc. European Conference on Computer Vision (ECCV)*, 2018.

[35] Stijn Vanderlooy and Eyke Hüllermeier. A critical analysis of variants of the auc. *Machine Learning*, 72(3):247–262, 2008.

[36] Ju-Chiang Wang, Yi-Hsuan Yang, I-Hong Jhuo, Yen-Yu Lin, Hsin-Min Wang, et al. The acousticvisual emotion guassians model for automatic generation of music video. In *Proc. ACM international conference on Multimedia*, pages 1379–1380, 2012.

[37] Shaomin Wu, Peter Flach, and Cèsar Ferri. An improved model selection heuristic for auc. In *Proc. European Conference on Machine Learning*, pages 478–489. Springer, 2007.

[38] Baohan Xu, Yanwei Fu, Yu-Gang Jiang, Boyang Li, and Leonid Sigal. Video emotion recognition with transferred deep feature encodings. In *Proc. ACM on International Conference on Multimedia Retrieval*, pages 15–22, 2016.

[39] Yi-Hsuan Yang and Homer H Chen. Machine recognition of music emotion: A review. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(3):40, 2012.

[40] Jong-Chul Yoon, In-Kwon Lee, and Siwoo Byun. Automated music video generation using multi-level feature-based segmentation. *Multimedia Tools and Applications*, 41(2):197, 2009.

[41] Yi Yu, Zhijie Shen, and Roger Zimmermann. Automatic music soundtrack generation for outdoor videos from contextual sensor information. In *Proc. ACM international conference on Multimedia*, pages 1377–1378, 2012.

[42] Yichi Zhang, Bryan Pardo, and Zhiyao Duan. Siamese style convolutional neural networks for sound search by vocal imitation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(2):429–441, 2019.

[43] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. *arXiv preprint arXiv:1804.03160*, 2018.

[44] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. Exploring principles-of-art features for image emotion recognition. In *Proc. ACM International Conference on Multimedia*, pages 47–56, 2014.