The Influence of Audio on Video Memorability with an Audio Gestalt Regulated Video Memorability System

Lorin Sweeney
Insight Centre for Data Analytics
Dublin City University
Dublin, Ireland
lorin.sweeney8@mail.dcu.ie

Graham Healy
Insight Centre for Data Analytics
Dublin City University
Dublin, Ireland
graham.healy@dcu.ie

Alan F. Smeaton
Insight Centre for Data Analytics
Dublin City University
Dublin, Ireland
alan.smeaton@dcu.ie

***Abstract*—Memories are the tethering threads that tie us to the world, and memorability is the measure of their tensile strength. The threads of memory are spun from fibres of many modalities, obscuring the contribution of a single fibre to a thread's overall tensile strength. Unfurling these fibres is the key to understanding the nature of their interaction, and how we can ultimately create more meaningful media content. In this paper, we examine the influence of audio on video recognition memorability, finding evidence to suggest that it can facilitate overall video recognition memorability rich in high-level (gestalt) audio features. We introduce a novel multimodal deep learning-based late-fusion system that uses audio gestalt to estimate the influence of a given video's audio on its overall short-term recognition memorability, and selectively leverages audio features to make a prediction accordingly. We benchmark our audio gestalt based system on the Memento10k short-term video memorability dataset, achieving top-2 state-of-the-art results.**

***Index Terms*—Memorability, multimodal, audio gestalt, deep learning**

I. Introduction and Related Work

Memories are the keepers of our continuity of self—without them, the very fabric of our being would fray. Yet, we scarcely have any influence on what we will ultimately remember or forget. The brain presides over the mechanisms of our memory from an opaque glass office, exercising sole editorial influence over its edifice. In fact, our odds of guessing what we will remember aren't much better than chance [1]. This lack of meta-cognitive insight, which prevents us from diving into our unconscious undercurrent, is what motivates and brings meaning to the exploration of the determinants of memory, and more specifically memorability—generally known as the likelihood that something will be remembered or forgotten. Naturally, the quantification of memorability is dependent on how we measure remembrance—which is in turn dependent on the modality and measurement paradigm.

Broadly speaking, there are two ways to measure remembrance: as *recognition*, where amidst content presentation, participants indicate which items they feel they have previously perceived; or as *recall*, where participants recount as much information as they can concerning previously presented content. These two measures respectively align with the two memory processes posited by the psychological dual process model of memory called *process dissociation* [2]. The first memory process is rapid, unconscious, and driven by a feeling of familiarity while the other is slower, conscious, and driven by a detail retrieving intention.

The three most common modalities with which memorability is explored are visual, textual, and auditory.

A. Visual Memorability

The predominant visual memorability measurement paradigm is *recognition*—where memorability is commonly defined as the percentage of correctly recognised targets [3]. Using this paradigm, a high degree of human consistency concerning which images are remembered or forgotten is observed, suggesting that "recognition memorability" is an intrinsic property of an image. Recognition memorability is also robust for other types of items, such as abstract visualisations [4], and specific objects within scenes [5]. This intrinsic property is not limited to static images, with faces shown to be consistently memorable across expression and viewpoint distortions [6], and videos shown to be highly consistent in memory performance for both soundless 10-second movie clips [7], and 3-6 second viral videos [8].

These results suggest that recognition memorability may be an intrinsic attribute of a wide range of stimulus types, even those with very different visual and semantic structures. Accordingly, relating it to other well-characterised image properties, such as saliency; colour features; aesthetics; etc., has been an active area of research. While several characteristics that correlate with recognition memorability have been proposed, a fully defining combination of features has yet to be identified. Simple image features, such as hue; saturation; or spatial frequency, have repeatedly been found not to correlate with recognition memorability [1], [5], [6]. The number of objects depicted in an image does not appear to directly

This work was partly supported by Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2, co-funded by the European Regional Development Fund.

relate to its overall recognition memorability [9], and likewise with properties such as aesthetics and interestingness [1]. However, combinations of semantically based attributes, such as object/scene category, emotion or actions, are predictive of recognition memorability [10]. Additionally, scrambled images retain consistencies in recognition memorability, but only for short time periods (seconds). These findings suggest that recognition memorability is more closely linked to high-level perceptual properties of an image rather than low-level visual properties.

B. Textual Memorability

Many basics visual recall memorability findings—recall as a function of serial presentation position—are also observed in the textual equivalent [11]. However, repeated recall has been found to incrementally increase subsequent recall performance for images, but not for words [12]. Words that arouse stronger emotion and are easier to visualise, exhibit enhanced recall [13], while concrete words that refer to things that can be experienced by the senses, have a relative advantage in recall over abstract words. Words with smaller sets of associated words have an advantage over those with larger sets [14]. Minimally counter-intuitive concepts have been found to lead to better recall, suggesting that recall memorability is not an inherent property of a concept, but a property of the concept in the context it is presented [15].

Similar to images, the recognition memorability of simple words is highly consistent across individuals, suggesting that it is an intrinsic property of words [16]. Less familiar, lower-frequency words [17]; imageable and concrete words [18], emotionally salient words [19] and the semantic context [20] in which they are presented, all enhance recognition memorability. Additionally, meanings of words are retained in favour of their lexical properties [21].

C. Auditory Memorability

Research into audio recall memorability shows that naming or verbalising sounds (phonological-articulation) can improve recall [22], and accordingly, non-verbal sounds have lower recall than verbal sounds [23]. Emotionality is known to play an important role in memory formation, and the emotional impact of a sound is correlated with the clarity of its perceived source [24]. Human activity is considered to be a positively valenced sound [25], and positive valence improves sound recall [26]. It is generally accepted that auditory recall memorability is inferior to visual recall memorability, and decays more quickly [27]. However, it is important not to overlook the role of the audio modality when exploring multi-modal media memorability, as multi-sensory experiences exhibit increased recall accuracy compared to uni-sensory ones [28], and sounds have the potential to provide valuable contextual priming information [29].

While little research has been conducted on auditory recognition memorability, interest has started to grow. Recent research suggests that similar to images and words, recognition memorability is an intrinsic property of sounds [30].

D. The MediaEval Memorability Task

The MediaEval2020 memorability task [8] is an annual event which benchmarks the effectiveness of predicting video memorability automatically. In 2020 this operated on video data which included audio for the first time, and several participants included audio features in their approaches to computing video memorability. Our approach and submission to the benchmark included audio gestalt features and produced promising preliminary results on the development test set of videos, shown in Table I. Due to the abnormally low, participant-wide results on the official validation set shown in Table II, and an omission in our official submissions, very little insight into the efficacy of our approach was gained. However, one of our audio-based submissions [31] did achieve the best-in-class results from among all participants for long-term memorability predictions.

TABLE I
MEDIAEVAL2020 MEDIA MEMORABILITY TASK RESULTS ON 200 DEV-SET VIDEOS KEPT FOR VALIDATION FOR EACH OF OUR RUNS.

	Short-term	Long-term
Run	**Spearman**	**Spearman**
Aug Captions + Spectrogram	0.345	0.365
Captions + Frames	0.338	0.437
Everything	0.319	0.425
Audio Gestalt Spectrogram	**0.364**	**0.470**
memento10k	0.314	-

TABLE II
OFFICIAL MEDIAEVAL2020 MEDIA MEMORABILITY TASK RESULTS ON TEST-SET FOR EACH OF OUR SUBMITTED RUNS.

	Short-term	Long-term
Run	**Spearman**	**Spearman**
Aug Captions + Spectrogram	0.054	**0.113**
Captions + Frames	0.05	0.059
Everything	-	0.119
Audio Gestalt Spectrogram	0.076	0.041
memento10k	**0.137**	-

In this paper, we evaluate the utility of including the audio modality in short-term video "recognition memorability" prediction, and assess our gestalt based video memorability prediction system by benchmarking it on the Memento10k dataset [32], comparing it to state-of-the-art solutions. Our contributions are two-fold: A) we assess the influence of the audio modality on video memorability B) we propose a multimodal deep learning-based late fusion system that uses audio gestalt to estimate the influence of the audio modality on overall video memorability, and selectively leverage audio features accordingly. Due to the nature of the Memento10k dataset, the recognition memorability in question is short-term.

II. METHODOLOGY

A. Audio Gestalt Regulated Video Memorability

Our system is a multimodal deep-learning based late fusion framework that uses an audio gestalt conditional mechanism to predict short-term video recognition memorability Figure 1.

Depending on an audio gestalt threshold (0.8), one of two pathways—*without audio*, using textual and visual features; and *with audio*, using textual, visual, and auditory features—is used to predict a video's recognition memorability score. The *without audio* stream's predictions are the weighted sum of our *Frame* model (0.38), and *Caption* model (0.62), while the *with audio* stream's predictions are the weighted sum of our *Frame* model (0.4), *Augmented Caption* model (0.47), and *Spectrogram* model (0.13). Both the weightings of the models's predictions and the gestalt threshold are determined using Randomised Search Cross-Validation (RSCV) from 0 to 1, in increments of 0.01.

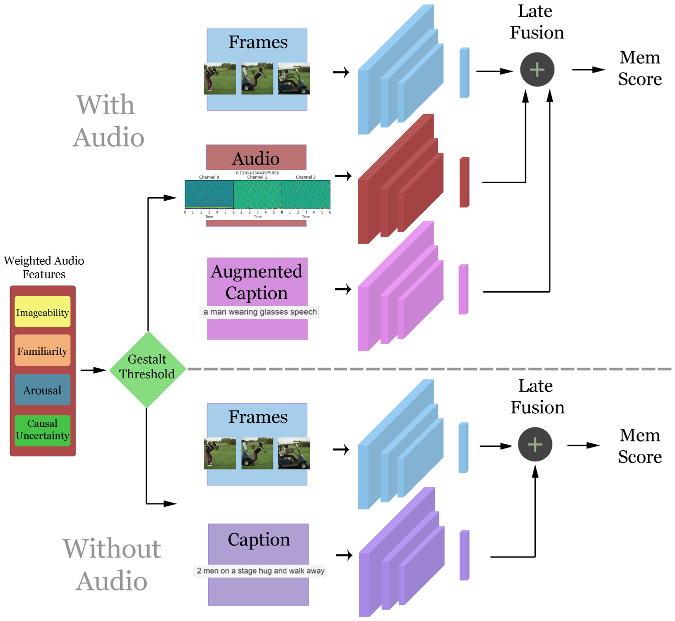

Fig. 1. Our multimodal deep-learning based late fusion framework, using a conditional audio gestalt based threshold.

B. Audio Gestalt

The Gestalt principles were first introduced by [33] in 1928, and continue to be relevant in modern psychology. Traditionally thought of as rules that characterise the organisation of visual scenes—helping us understand them better—the Gestalt principles of *similarity*; *connectedness*; *common region*; *spatial proximity* [34], and *goodness* [35] have been shown to benefit visual recognition memorability.

The very first usage of the term Gestalt was in 1890 in [36], which observed that humans can recognise two identical melodies even when no two corresponding notes have the same frequency. It was suggested that this property indicated the presence of a "Gestalt quality"—a conceptual characteristic that assists our "big picture" understanding of complex sensory data composed of many different parts. Unfortunately, since then, few insights intersecting audio gestalt and other well established audio properties have been revealed. The concept of gestalt in the context of audio was recently reintroduced by [30], using the term gestalt to encapsulate high-level conceptual audio features. They found the following gestalt features: imageability; human causal uncertainty (Hcu); arousal; and familiarity, to be strongly correlated with audio memorability. In in this paper, we aim to practically apply these findings with the goal of elucidating the role of audio in overall video recognition memorability.

We create our own audio gestalt predictor using a weighted sum of our proxy measures for these four features. RSCV between 0 and 1 in increments of 0.05 is used to determine each of the weights. Due to the strong negative correlation between sound imageability and musicality [37], we predicate imageability on whether the audio is classified as music or not. We use the PANNs [38] network to generate audio-tags, labelling the audio as music (giving it a score of 1.0) if a musical tag is present in the top 75% confidence. Hcu and arousal scores are independently predicted with ImageNet-pretrained xResNet34 models fine-tuned on spectrograms from the HCU400 dataset [24]. Due limited available options, for familiarity, we use the top audio-tag confidence score of the PANNs [38] network as a proxy (Spearman = 0.305, pval = 4.749e-10 between the two scores in the HCU400 dataset). These four scores are then normalised (scaled into a 0-1 range), and a weighted score (with weights of 0.2, 0.2, 0.2, and 0.4 respectively) is calculated to produce an audio gestalt score.

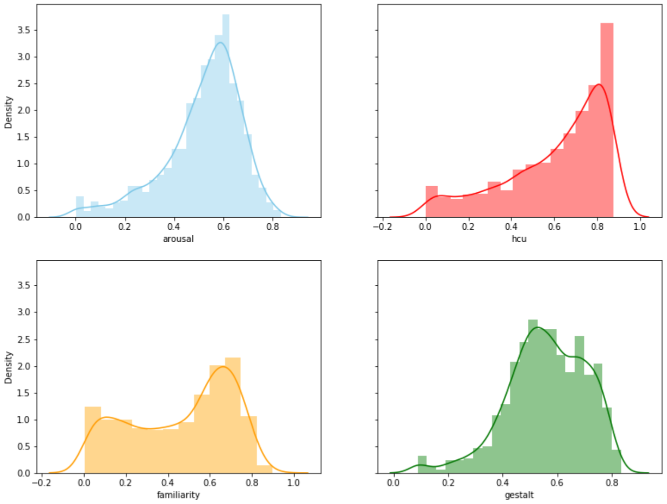

Fig. 2. Distribution of audio gestalt and gestalt related audio features from 1,468 validation videos.

C. Auditory Features

For auditory features, we train a network to predict a video's recognition memorability from audio spectrograms—our *Spectrogram model*. We extracted Mel-frequency cepstral coefficients (n_fft:2048, hop_length:256, n_mels:128) from the the 6,890 Memento10k [32] training videos with audio, and stacked them with their delta coefficients in order to create three channel spectrogram images. These spectrogram images are then used to train an ImageNet-pretrained xResNet34 model for 15 epochs; with a max learning rate of 1e-2; and weight decay of 1e-3 to predict audio recognition

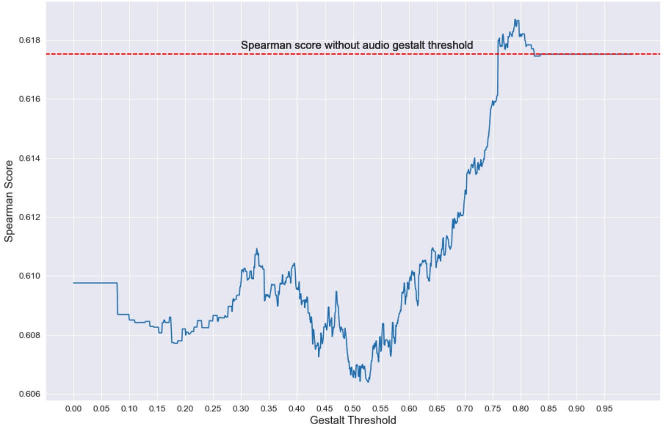

Fig. 3. Effect of gestalt thresholds on Spearman scores of 1,468 Memento10k validation videos.

memorability. Additionally, we fit a Bayesian Ridge Regressor with VGGish [39] audio features—our *Bayesian Ridge* model. We extract 128-dimensional embeddings for each second of video audio, resulting in a 384-dimensional feature set per video.

D. Visual Features

We evaluate the extent to which static visual features contribute to video recognition memorability by training a network to predict a video's recognition memorability from a single frame—our *Frame* model. We train an ImageNet-pretrained xResNet50 to predict image recognition memorability by first training on the LaMem dataset [3] for 50 epochs; with a maximum learning rate of 3e-2; and weight decay of 1e-2, and then fine-tuning on the 6,890 Memento10k [32] training videos which have audio with the same hyperparameters. At test time, a video's recognition memorability score is calculated by averaging predictions of the first, middle, and last frame.

E. Textual Features

For textual features, we train a network to predict a video's recognition memorability from a paragraph of text composed of five independently generated human captions—our *Caption model*. Given that overfitting is a primary concern, we use the AWD-LSTM (ASGD Weight-Dropped LSTM) architecture [40], as it is highly regularised, and is comparable to other state-of-the-art language models. In order to fully take advantage of the high level representations that a language model offers, we transfer train our model using UMLFiT [41], a method that uses discriminative fine-tuning, slanted triangular learning rates, and gradual unfreezing to avoid catastrophic forgetting.

A Wiki-103-pretrained language model is fine-tuned on the first 300,000 captions from Google's Conceptual Captions dataset [42] for a total of 10 epochs with a dropout multiplier of 0.5 and max learning rate of 2e-3, resulting in a final language model accuracy of 37%. The encoder from that model is re-used in another model of the same architecture, but trained on captions from the 6,890 Memento10k [32] training videos with audio, for a total of 15 epochs with a dropout multiplier of 0.8 and a max learning rate of 1e-3, to predict recognition memorability scores, rather than the next word in a sentence. An additional network is trained the same way, but fine-tuned on captions that are augmented with audio tags extracted using the PANNs [38] network—our *Augmented Caption model*.

In all cases, models were independently trained on the 6,890 Memento10k training set videos with audio, and independently validated on the 1,484 Memento10k validation videos with audio. All parameter tuning (e.g. RSCV) was performed using the Memento10K training set.

III. Results

As in the MediaEval memorability task, prediction performance is measured by calculating the Spearman's rank correlation of the predicted memorability rankings with their ground truth rankings. Table III shows the Spearman rank correlation scores of the individual components of our audio gestalt system, many of their combinations, and the final implementation of our audio gestalt system on the 1,484 Memento10k validation videos with audio. The best performing individual component is our *Caption model*, achieving a Spearman score of 0.5710. Each of the component combinations are the result of a randomised search weighted summation of their predictions, with the best combination being *Captions* + *Frames* (0.6175). Our audio gestalt based system was the best performing approach, achieving a Spearman score of 0.6181.

TABLE III
Results on 1,484 Memento10k validation videos with audio.

	Memorability
Approach	**Spearman**
Spectrogram	0.2030
Bayesian Ridge	0.2913
Frames	0.4808
Frames + Spectrogram	0.4876
Frames + Bayesian Ridge	0.4992
Captions	0.5710
Captions + Spectrogram	0.5715
Captions + Bayesian Ridge	0.5741
Augmented Captions	0.5555
Augmented Captions + Spectrogram	0.5562
Augmented Captions + Bayesian Ridge	0.5576
Augmented Captions + Frames	0.6068
Captions + Frames	0.6175
Everything Ridge	0.6066
Everything Spectrogram	0.6061
Audio Gestalt Ridge Normal Captions	0.6175
Audio Gestalt Spectrogram Normal Captions	0.6176
Audio Gestalt Ridge	0.6181
Audio Gestalt Spectrogram	**0.6181**

To evaluate the effectiveness of our approach, we compare against the Memento10k benchmark scores [32]. From Table IV we can see that our audio gestalt based approach outperforms all other approaches except SemanticMemNet [32]—the model introduced alongside the Memento10k dataset.

TABLE IV

Comparison of state-of-the-art on Memento10k. *Trained and validated on fewer videos due to audio constraint, 7,000 vs. 6,890 and 1,500 vs. 1484 respectively.

	Memorability
Approach	**Spearman**
Human Consistency	0.730
MemNet Baseline [3]	0.485
Cohendet et al. (Semantic) [10]	0.552
Cohendet et al. (ResNet3D) [10]	0.574
Feature Extraction + Regression (as in [43])	0.615
SemanticMemNet [32]	0.663
Audio Gestalt	**0.618***

With respect to our results in Table III, the general trend for predicting video recognition memorability seems to be that the more modalities used, the better the predictions. Even the addition of a poorly-performing individual audio model (0.2913) with a better-performing individual visual model (0.4808), produces an increase in performance (0.4992). There are however, some very important exceptions to this trend. Indiscriminately tri-modal approaches, *Everything Ridge* (0.6066) and *Everything Spectrogram* (0.6061), achieve lower Spearman scores than the bi-modal combination of visual and textual predictions (0.6175), and their selectively tri-modal counterparts (0.6181).

At first glace, it appears that augmenting captions with audio-tags is worse than vanilla captions, Augmented Captions (0.5555) vs Captions (0.5710); Augmented Captions + Spectrogram (0.5562) vs Captions + Spectrogram (0.5715); Augmented Captions + Bayesian Ridge (0.5576) vs Captions + Bayesian Ridge (0.5741); Augmented Captions + Frames (0.6068) vs Captions + Frames (0.6175), however, when selectively used in our audio gestalt system (0.6181), they outperform vanilla captions (0.6175).

IV. Discussion

Our audio gestalt based system ultimately outperforms all of our other tested approaches. Even though the advantage incurred is only marginal, selectively including audio features (0.6181) is ultimately better than both always including them (0.6066), and not including them (0.6175). We believe that this can in part be explained by the fact that sounds have the have the potential to provide valuable contextual priming information [29], but that some sounds simply add noise, having a deleterious effect on overall understanding of a context. Thinking of audio gestalt as an ontological property that encapsulates high-level auditory features that positively contribute towards our understanding of a context, helps explain the benefit of using it as a measure to discriminate between useful and distracting audio in multimodal content. The effect of different gestalt thresholds is shown in Figure 3.

It is interesting to note that there is no difference in Spearman score between Audio Gestalt Spectrogram (0.6181) and Audio Gestalt Ridge (0.6181), even though the Bayesian Ridge achieves a noticeably higher Spearman score (0.2913) than the Spectrogram model (0.2030). This indicates that the inclusion of auditory features is not strictly additive, and further suggests that they may act as a contextual signal of some sort.

In [30], they found that the strongest predictors of sound recognition memorability were imageability, and causal uncertainty (Hcu). Naturally, we would expect our audio gestalt weightings to reflect this to some degree, but we found that the highest weighted audio gestalt feature is familiarity (top audio-tag confidence score). The gestalt weightings for imageability; Hcu; familiarity; and arousal, are 0.2; 0.2; 0.4; 0.2 respectively. As shown in Figure 2, familiarity is the only audio gestalt feature with a bi-modal distribution. Both arousal and Hcu are heavily left skewed, leading us to believe that the models used to predict their scores have been overfit, and can be improved.

V. Conclusions

In this paper we have assessed the influence of the audio modality on video recognition memorability, finding evidence to suggest that it primarily plays a contextualising role, with the potential to act as a signal or trigger that aids recognition depending on the extent of its high-level features. We introduced a novel multimodal deep learning-based late-fusion system that uses audio gestalt to estimate the influence of a given video's audio on its overall short-term recognition memorability that selectively leverages audio features to make a prediction accordingly. Our findings add further credibility to the hypothesis that recognition memorability is more closely linked to high-level perceptual properties of content than low-level properties, and that this relationship extends beyond the visual domain. Similar to the way in which textual memorability, both recall and recognition, has been suggested as not an inherent property of a concept, but a property of the concept in the context it is presented, the influence of auditory memorability in a multimodal medium such as video, is likely to be highly context dependent.

While this work has made progress towards understanding the influence of the audio modality on short-term video recognition memorability, the full extent of its role is far from being understood. It is possible that the correlation between the content/context of a video's auditory modality and its visual modality could play an important role in determining the audio's impact on the video's overall recognition memorability, however, without testing this experimentally, we simply cannot answer this question. We believe that improvements can be made by refining our measure of audio gestalt.

Independent memorability scores for each of the modalities—audio, visual, and textual—would assist us in elucidating the role they each play when coinciding with one another in a multimodal medium such as video, and should be a focus of future memorability research. Similarly, recognition memorability and recall memorability would each benefit from a directed disentanglement effort as they are often conflated. The way in which they interact is relatively unexplored, and further study here is likely to yield valuable insights.

References

[1] P. Isola, J. Xiao, D. Parikh, A. Torralba, and A. Oliva, "What makes a photograph memorable," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 7, pp. 1469–1482, 2013.

[2] L. Jacoby, "A process dissociation framework: Separating automatic from intentional uses of memory," *Journal of Memory and Language*, vol. 30, no. 5, pp. 513–541, 1991.

[3] A. Khosla, A. S. Raju, A. Torralba, and A. Oliva, "Understanding and predicting image memorability at a large scale," in *Proc. IEEE International Conference on Computer Vision*, 2015, pp. 2390–2398.

[4] M. A. Borkin, A. A. Vo, Z. Bylinskii, P. Isola, S. Sunkavalli, A. Oliva, and H. Pfister, "What makes a visualization memorable?" *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2306–2315, 2013.

[5] R. Dubey, J. Peterson, A. Khosla, M.-H. Yang, and B. Ghanem, "What makes an object memorable?" in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1089–1097.

[6] W. A. Bainbridge, "The memorability of people: Intrinsic memorability across transformations of a person's face." *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 43, no. 5, p. 706, 2017.

[7] R. Cohendet, K. Yadati, N. Q. Duong, and C.-H. Demarty, "Annotating, understanding, and predicting long-term video memorability," in *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, 2018, pp. 178–186.

[8] A. García Seco de Herrera, R. Savran Kiziltepe, J. Chamberlain, M. G. Constantin, C.-H. Demarty, F. Doctor, B. Ionescu, and A. F. Smeaton, "Overview of MediaEval 2020 predicting media memorability task: What makes a video memorable?" in *Working Notes Proceedings of the MediaEval 2020 Workshop*, 2020.

[9] P. Isola, J. Xiao, A. Torralba, and A. Oliva, "What makes an image memorable," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 145–152.

[10] R. Cohendet, C.-H. Demarty, N. Q. Duong, and M. Engilberge, "Videomem: Constructing, analyzing, predicting short-term and long-term video memorability," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 2531–2540.

[11] J. Deese and R. A. Kaufman, "Serial effects in recall of unorganized and sequentially organized verbal material." *Journal of Experimental Psychology*, vol. 54, no. 3, p. 180, 1957.

[12] M. H. Erdelyi and J. Becker, "Hypermnesia for pictures: Incremental memory for pictures but not words in multiple recall trials," *Cognitive Psychology*, vol. 6, no. 1, pp. 159–171, 1974.

[13] M. Bock, "The influence of emotional meaning on the recall of words processed for form or self-reference," *Psychological Research*, vol. 48, no. 2, pp. 107–112, 1986.

[14] D. L. Nelson and T. A. Schreiber, "Word concreteness and word structure as independent determinants of recall," *Journal of Memory and Language*, vol. 31, no. 2, pp. 237–260, 1992.

[15] M. A. Upala, L. O. Gonce, R. D. Tweney, and D. J. Slone, "Contextualizing counterintuitiveness: How context affects comprehension and memorability of counterintuitive concepts," *Cognitive Science*, vol. 31, no. 3, pp. 415–439, 2007.

[16] K. Mahowald, P. Isola, E. Fedorenko, E. Gibson, and A. Oliva, "Memorable words are monogamous: The role of synonymy and homonymy in word recognition memory," PsyArXiv, 2018.

[17] V. M. Garlock, A. C. Walley, and J. L. Metsala, "Age-of-acquisition, word frequency, and neighborhood density effects on spoken word recognition by children and adults," *Journal of Memory and Language*, vol. 45, no. 3, pp. 468–492, 2001.

[18] P. Klaver, J. Fell, T. Dietl, S. Schür, C. Schaller, C. E. Elger, and G. Fernández, "Word imageability affects the hippocampus in recognition memory," *Hippocampus*, vol. 15, no. 6, pp. 704–712, 2005.

[19] E. A. Kensinger and S. Corkin, "Memory enhancement for emotional words: Are emotional words more vividly remembered than neutral words?" *Memory & Cognition*, vol. 31, no. 8, pp. 1169–1180, 2003.

[20] L. L. Jacoby and M. Dallas, "On the relationship between autobiographical memory and perceptual learning." *Journal of Experimental Psychology: General*, vol. 110, no. 3, p. 306, 1981.

[21] I. Begg and W. A. Wickelgren, "Retention functions for syntactic and lexical vs semantic information in sentence recognition memory," *Memory & Cognition*, vol. 2, no. 2, pp. 353–359, 1974.

[22] J. C. Bartlett, "Remembering environmental sounds: The role of verbalization at input," *Memory & Cognition*, vol. 5, no. 4, pp. 404–414, 1977.

[23] A. Paivio, R. Philipchalk, and E. J. Rowe, "Free and serial recall of pictures, sounds, and words," *Memory & Cognition*, vol. 3, no. 6, pp. 586–590, 1975.

[24] I. Ananthabhotla, D. B. Ramsay, and J. A. Paradiso, "Hcu400: An annotated dataset for exploring aural phenomenology through causal uncertainty," in *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 920–924.

[25] D. Dubois, C. Guastavino, and M. Raimbault, "A cognitive approach to urban soundscapes: Using verbal data to access everyday life auditory categories," *Acta acustica united with acustica*, vol. 92, no. 6, pp. 865–874, 2006.

[26] L. Jäncke, "Music, memory and emotion," *Journal of biology*, vol. 7, no. 6, pp. 1–5, 2008.

[27] J. Bigelow and A. Poremba, "Achilles' ear? inferior human short-term and recognition memory in the auditory modality," *PloS One*, vol. 9, no. 2, p. e89914, 2014.

[28] A. Thelen, D. Talsma, and M. M. Murray, "Single-trial multisensory memories affect later auditory and visual object discrimination," *Cognition*, vol. 138, pp. 148–160, 2015.

[29] A. Schirmer, Y. H. Soh, T. B. Penney, and L. Wyse, "Perceptual and conceptual priming of environmental sounds," *Journal of Cognitive Neuroscience*, vol. 23, no. 11, pp. 3241–3253, 2011.

[30] D. Ramsay, I. Ananthabhotla, and J. Paradiso, "The intrinsic memorability of everyday sounds," in *Audio Engineering Society Conference: 2019 AES Intnl. Conference on Immersive and Interactive Audio*, 2019.

[31] L. Sweeney, G. Healy, and A. F. Smeaton, "Leveraging audio gestalt to predict media memorability," in *MediaEval Multimedia Benchmark Workshop Working Notes, arXiv preprint arXiv:2012.15635*, 2020.

[32] A. Newman, C. Fosco, V. Casser, A. Lee, B. McNamara, and A. Oliva, "Multimodal memorability: Modeling effects of semantics and decay on video memorability," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 223–240.

[33] M. Wertheimer, "Laws of organization in perceptual forms," in *A source book of Gestalt psychology*, W. Ellis, Ed. Kegan Paul, Trench, Trubner & Company, 1938, ch. 5, p. 71–88.

[34] D. J. Peterson and M. E. Berryhill, "The gestalt principle of similarity benefits visual working memory," *Psychonomic Bulletin & Review*, vol. 20, no. 6, pp. 1282–1289, 2013.

[35] L. Goetschalckx, P. Moors, S. Vanmarcke, and J. Wagemans, "Get the picture? goodness of image organization contributes to image memorability," *Journal of Cognition*, vol. 2, no. 1, 2019.

[36] C. v. Ehrenfels, "Über gestaltqualitäten," *Vierteljahrsschrift für wissenschaftliche Philosophie*, vol. 14, no. 3, pp. 249–292, 1890.

[37] A. R. Bowles, C. B. Chang, and V. P. Karuzis, "Pitch ability as an aptitude for tone learning," *Language Learning*, vol. 66, no. 4, pp. 774–808, 2016.

[38] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "PANNs: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880–2894, 2020.

[39] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, "Cnn architectures for large-scale audio classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 131–135.

[40] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and optimizing LSTM language models," in *International Conference on Learning Representations*, 2018.

[41] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 328–339.

[42] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proceedings 56th Annual Meeting of the Association for Computational Linguistics (Vol 1: Long Papers)*, 2018, pp. 2556–2565.

[43] S. Shekhar, D. Singal, H. Singh, M. Kedia, and A. Shetty, "Show and recall: Learning what makes videos memorable," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2017, pp. 2730–2739.