Coordinated Joint Multimodal Embeddings for Generalized Audio-Visual Zero-shot Classification and Retrieval of Videos

Kranti Kumar Parida
IIT Kanpur

Neeraj Matiyali
IIT Kanpur

Tanaya Guha
University of Warwick

Gaurav Sharma
NEC Labs America

Abstract

We present an audio-visual multimodal approach for the task of zero-shot learning (ZSL) for classification and retrieval of videos. ZSL has been studied extensively in the recent past but has primarily been limited to visual modality and to images. We demonstrate that both audio and visual modalities are important for ZSL for videos. Since a dataset to study the task is currently not available, we also construct an appropriate multimodal dataset with 33 classes containing 156, 416 videos, from an existing large scale audio event dataset. We empirically show that the performance improves by adding audio modality for both tasks of zero-shot classification and retrieval, when using multimodal extensions of embedding learning methods. We also propose a novel method to predict the 'dominant' modality using a jointly learned modality attention network. We learn the attention in a semi-supervised setting and thus do not require any additional explicit labelling for the modalities. We provide qualitative validation of the modality specific attention, which also successfully generalizes to unseen test classes.

1. Introduction

Zero-shot learning (ZSL) refers to the setting when test time data comes from classes that were not seen during training. In the past few years, ZSL for classification has received significant attention [1–9] due to the challenging nature of the problem, and its relevance to real world settings, where a trained model deployed in the field may encounter classes for which no examples were available during training. Initially, ZSL was proposed and studied in the setting where the test examples were from unseen classes and were classified into one of the unseen classes only [6]. This however is an artificial/controlled setting. More recent ZSL works thus focus on a setting where unseen test examples are classified into both seen and unseen classes [2, 9, 10]. The present work follows the latter setting known as the *Generalized ZSL.*

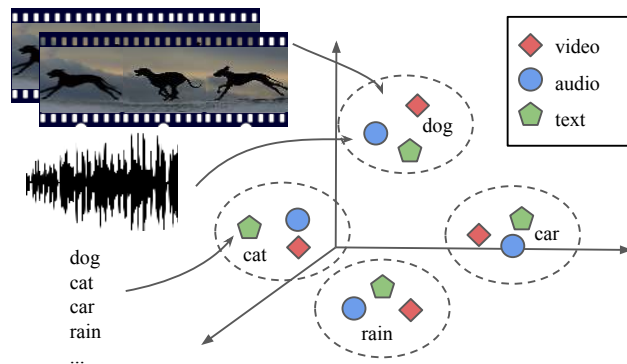The majority of work involving generalized ZSL [3, 10]

Figure 1. Illustration of the proposed method. We jointly embed all videos, audios and text labels into the same embedding space. We learn the space such that the corresponding embedding vectors for the same classes have lower distances than those of different classes. Once embeddings are learned, ZSL classification and crossmodal retrieval can be posed as a nearest neighbor search in the embedding space.

has (i) worked with images, and (ii) used only visual representations along with text embeddings of the classes. When dealing with images, this is optimal. However, for the task of video ZSL, the audio modality, if available, may help with the task by providing complementary information. Ignoring the audio modality completely might even render an otherwise easy classification task difficult, eg. if we are looking to classify an example from the 'dog' class, the dog might be highly occluded and not properly visible in the video, but the barking sound might be prominent.

In this work, we study the problem of ZSL for videos with general classes like 'dog', 'sewing machine', 'ambulance', 'camera', 'rain', and propose to use audio modality in addition to the visual modality. ZSL for videos is relatively less studied, cf. ZSL for images. There are several works on video ZSL for the specific task of human action recognition [11–13] but they ignored the audio modality as well. Our focus here is on leveraging both audio and video modalities to learn a joint projection space for audio, video and text (class labels). In such an embedding space, ZSL tasks can be formulated as nearest neighbor searches (Fig. 1

illustrates the point). When doing classification, a new test video is embedded into the space and the nearest class embedding is predicted to be its class. Similarly, when doing retrieval, the nearest video or audio embeddings are predicted to be its semantic retrieval outputs.

We propose cross-modal extensions of the embedding-based ZSL approach based on triplet loss for learning such a joint embedding space. We optimize an objective based on (i) two cross-modal triplet losses, one each for ensuring compatibility between the text (class labels) and the video, and the text and the audio, and (ii) another loss based on crossmodal compatibility of the audio and visual embeddings. While the triplet losses encourage the audio and video embeddings to come closer to respective class embeddings in the common space, the audio-visual crossmodal loss encourages the audio and video embeddings from the same sample to be similar. These losses together ensure that the three embeddings of the same class are closer to each other relative to their distance from those of different classes. The crossmodal loss term is an ℓ_2 loss, and uses paired audio-video data, the annotation being trivially available from the videos. While the text-audio and text-video triplet losses use class annotations available for the seen classes during training, the crossmodal term uses the trivial constraint that audio and video from the same example are similar.

As another contribution, we also propose a modality attention based extension, which first seeks to identify the 'dominant' modality and then makes a decision based on that modality only if possible. To clarify our intuition of 'dominant', we refer back to the dog video example above, where the dog may be occluded but barking is prominent. In this case, we would like the audio modality to be predicted as dominant, and subsequently be used to make the class prediction. In case the attention network is not able to decide a clear dominant modality the inference then continues using both the modalities. This leads to a more interpretable model which can also indicate which modality it is basing its decision on. Furthermore, we show empirically that using such attention learning improves the performance, and brings it to be competitive to model trained on a concatenation of both modality features.

A suitable dataset was not available for the task of audiovisual ZSL. Hence, we construct a multimodal dataset with class level annotations. The dataset is a subset of a recently published large scale dataset, called Audioset [14], which was primarily created for audio event detection and maintains a comprehensive sound vocabulary. We subsample the dataset to allow studying the task of audiovisual ZSL in a controlled setup. In particular, the subsampling ensures that (i) the classes have relatively high number of examples, with the minimum number of examples in any class being 292, (ii) the classes belong to diverse groups, eg. animals, vehi-cles, weather events, (ii) the set of unseen classes is such that the pre-trained video networks could be used without violating the zero-shot condition, ie. the pre-training did not involve classes close to the unseen classes in our dataset. We provide more details in Sec. 4.

In summary, our contributions are as follows. (i) We introduce the problem of audiovisual ZSL for videos, (ii) we construct a suitable dataset to study the task, (iii) we propose a multimodal embedding based ZSL method for classification and crossmodal retrieval, (iv) we propose a modality attention based method, which indicates which modality is dominant and was used to make the decision. We thoroughly evaluate our method on the dataset and show that considering audio modality, whenever appropriate, helps video ZSL tasks. We also show our method on standard ZSL datasets and results for some existing ZSL approaches for single-modality in our dataset as well. We also present qualitative results highlighting the improved cases using the proposed methods.

2. Related Work

Zero-shot learning (ZSL). ZSL has been quite popular for image classification [1–9, 15–18], and recently has been used for object detection in images as well [19–21]. The problem has been often addressed as a task of embedding learning, where the images and their class labels are embedded in a common space. The two types of class embeddings commonly used in the literature are based on (i) attributes like shape, color, and pose [2, 5, 6, 9], and (ii) semantic word embeddings [2, 4, 7, 9]. Few works have also used both the embeddings together [1, 8, 16]. Different from embedding learning, few recent works [2, 9, 22, 23] have proposed to generate the data for the unseen class using a generative approach conditioned on the attribute vectors. The classifiers are then learned using the original data for the seen classes and the generated data for the unseen classes. This line of work follows the recent success of image generation methods [24, 25]. The initially-studied setting in ZSL refers to the one where the test examples were classified into unseen test classes only [6]. However, more recently the generalized version was proposed where they are classified into both seen and unseen classes [3]. We address this later more practical setting[1].

Work on ZSL involving audio modality is scarce. We are aware of only one very recent work, where the idea of ZSL has been used to recognize unseen phonemes for multilingual speech recognition [26].

Audiovisual learning. In the last few years, there has been a significant growth in research efforts that leverage information from audio modality to aid visual learning tasks and

---

[1]Some earlier video retrieval works were called zero-shot, however, they are not strictly zero-shot in the current sense. Kindly see Supplementary material for a detailed discussion

vice-versa. Audio modality has been exploited for applications such as, audiovisual correspondence learning [27–30], audiovisual source separation [31, 32] and source localization [33–35]. Among the representative works, Owens et al. [28] used CNNs to predict, in a self-supervised way, if a given pair of audio and video clip is temporally aligned or not. The learned representations are subsequently used to perform sound source localization, and audio-visual action recognition. In a task of crossmodal biometric matching, Nagrani et al. [36] proposed to match a given voice sample against two or more faces. Arandjelovic et al. [33] introduced the task of audio-visual correspondence learning, where a network comprising visual and audio subnetworks was trained to learn semantic correspondence between audio and visual data. Along the similar lines, Arandjelovic et al. [29] and Sencoak et al. [35] investigated the problem of localizing objects in an image corresponding to a sound input. Gao et al. [32] proposed a multi-instance multilabel learning framework to address the audiovisual source separation problem, where they extract different audio components and associate them with the visual objects in a video. Ephrat et al. [37] proposed a joint audiovisual model to address the classical cocktail party problem (blind speech source separation). Zhao et al. [31] proposed a self-supervised learning framework to address the problem of pixel-level (audio) source localization [38].

3. Coordinated Joint Multimodal Embeddings

We now present our method in detail. Fig. 1 illustrates the basic idea and Fig. 2 gives the high level block diagram of the proposed method. Our method works by projecting all three inputs, audio, video and text, onto a common embedding space such that class constraints and crossmodal similarity constraints are satisfied. The class constraints are enforced using bimodal triplet losses between audio and text, and video and text embeddings. Denoting $\mathbf{a}_i, \mathbf{v}_i, \mathbf{t}_i$ as the audio, video and text embedding (we explain how we obtain them shortly) for an example i, we define the bimodal triplet losses as follows

$$\mathcal{L}_{TA}(\mathbf{a}_p, \mathbf{t}_p, \mathbf{a}_q, \mathbf{t}_q) = [d(\mathbf{a}_p, \mathbf{t}_p) - d(\mathbf{a}_q, \mathbf{t}_p) + \delta]_+ \quad (1)$$

$$\mathcal{L}_{TV}(\mathbf{v}_p, \mathbf{t}_p, \mathbf{v}_q, \mathbf{t}_q) = [d(\mathbf{v}_p, \mathbf{t}_p) - d(\mathbf{v}_q, \mathbf{t}_p) + \delta]_+ \quad (2)$$

where, $(\mathbf{a}_p, \mathbf{v}_p, \mathbf{t}_p)$ and $(\mathbf{a}_q, \mathbf{v}_q, \mathbf{t}_q)$ are two example videos with $\mathbf{t}_p \neq \mathbf{t}_q$ and both $\mathbf{t}_p, \mathbf{t}_q$ belong to a seen class each. These losses force the audio and video embeddings to be closer to the correct class embedding by a margin $\delta > 0$ cf. the incorrect class embeddings.

We also use a third loss to ensure the crossmodal similarity between the audio-video streams that come from the same video in the common embedding space. This loss is simply a ℓ_2 loss given by

$$\mathcal{L}_{AV}(\mathbf{a}_p, \mathbf{v}_p) = \|\mathbf{a}_p - \mathbf{v}_p\|_2^2. \quad (3)$$
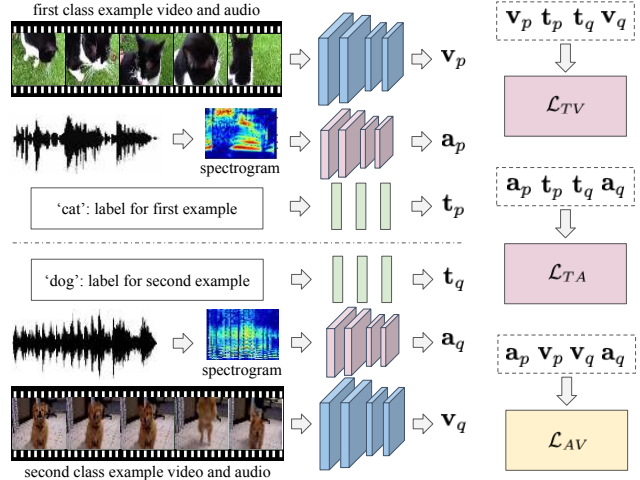

Figure 2. Block diagram of the proposed approach. Pairs of video, audio and text networks share weights.

The full loss function is thus a weighted average of these three losses.

$$\mathcal{L} = \lambda \sum_{p \in \mathcal{T}} \mathcal{L}_{AV} + \gamma \sum_{\substack{p,q \in \mathcal{T} \\ y_p \neq y_q}} \{\alpha_v \mathcal{L}_{TV} + \alpha_a \mathcal{L}_{TA}\}, \quad (4)$$

where, $\lambda, \gamma, \alpha_v, \alpha_a$ are the hyperparameters that control the contributions of the different terms, and $\mathcal{T}$ is the index set over the training examples $\{(\mathbf{a}_i, \mathbf{v}_i, y_i)|i = 1, \ldots, N\}$ with y_i being the class label. With these three losses over all pairwise combinations of the modalities, ie. $\mathcal{L}_{TV}, \mathcal{L}_{TA}, \mathcal{L}_{AV}$, we force the embeddings from all the three modalities to respect the class memberships and similarities.

Representations and parameters. We now need to specify the parameters over which these losses are optimized. We represent each of the three types of inputs, ie. audio, video, and text, using the corresponding state-of-the-art neural networks outputs which we denote as $f_a(\cdot), f_v(\cdot), f_t(\cdot)$. We project each representation with corresponding neural networks which are small MLPs, denoted as $g_a(\cdot), g_v(\cdot), g_t(\cdot)$ with parameters $\theta_a, \theta_v, \theta_t$ (we give details about all these networks in the implementation details Sec. 5). Finally, the representations are obtained by passing the input audio/video/text through the corresponding networks sequentially, ie. $\mathbf{x} = g_x \circ f_x(X)$ where $x \in \mathbf{a}, \mathbf{v}, \mathbf{t}$ and X is the corresponding raw audio/video/text input. We keep the initial network parameters fixed to be that of the pretrained networks and optimize over the parameters of the projection networks. Hence, the full optimization is given as,

$$\theta_a^*, \theta_v^*, \theta_t^* = \underset{\theta_a, \theta_v, \theta_t}{\arg \min} \mathcal{L}(\mathcal{T}). \quad (5)$$

We train for the parameters using standard backpropagation for neural networks.

Inference. Once the model has been learned, we use nearest neighbor in the embedding space for making predictions. In the case of classification, the audio and video are embedded in the space and the class embedding with the minimum average distance with them is taken as the prediction, ie.

$$\mathbf{t}^* = \arg\min_{\mathbf{t}} \left\{ d(\mathbf{a}, \mathbf{t}) + d(\mathbf{v}, \mathbf{t}) \right\}. \quad (6)$$

In the case of (crossmodal) retrieval, the sorted list of audio or video examples are returned as the result, based on their distance from the query in the embedding space.

Modality attention based learning. In the prequel, the method learns to make a prediction (classification or retrieval) using both the audio as well as video modalities. We augment our method to predict modality attention to find the dominant modality for each sample, eg. in case when the object is occluded or not visible, but the characteristic sound is clearly present we want the network to be able to make the decision based on the audio modality only. We incorporate such attention by adding an attention predictor network $f_{attn}(\cdot)$, with parameters θ_{attn}, which takes the concatenated audio and video features as inputs and predicts a scalar $\alpha \in [0, 1]$ which gives us the relative importance weights $\alpha_v = \alpha, \alpha_a = 1 - \alpha$ in eq. 4. All the network parameters are then learned jointly.

To further guide the attention network, we use the intuition that when one modality is dominant, say audio, the correct class embedding is expected to be much closer to the audio embedding, than the other classes cf. the video embedding. Hence the entropy of the prediction probability distribution over classes, for the dominant modality, should be very low. To compute such distribution, we first compute the inverse of the distances of the query embedding to all the class embeddings, and then ℓ_1 normalize the vector. We then derive a supervisory signal for α using the entropies computed w.r.t. audio and video modalities, denoted $e_a, e_v \in [0, \log N_c]$ where N_c is the number of classes over which prediction is being done, as

$$\alpha = \begin{cases} 0, & \text{if } e_v < e_a - \xi \\ 1, & \text{if } e_a < e_v - \xi \\ 0.5, & \text{otherwise} \end{cases} \quad (7)$$

where, $\xi > 0$ is a threshold parameter for preferring one of the modalities based on their entropy difference. The modality attention objective becomes, $\mathcal{L}_{attn} = \mathcal{L} + \mathcal{L}_{CE-\alpha}$, where $\mathcal{L}$ is the objective from eq. 4, $\mathcal{L}_{CE-\alpha}$ is the cross entropy loss on α based on the generated supervision above. This loss is minimized jointly over all $\theta_a, \theta_v, \theta_t, \theta_{attn}$.

Modality selective inference with attention. While attention is interesting at training as it helps identify the dominant modality and learn better models. We also use attention to make inference using only the predicted dominant modality at test time. When the predicted attention is higher than a threshold for one of the modalities we only compute distance for that modality in the embedding space and use that to make the prediction.

We could also use the above computed α value based on the difference of entropies of the prediction distributions (eq. 7) at test time, even when not training with modality attention. We use that as a baseline to verify that learning to predict the attention helps improve the performance.

Calibrated stacking in generalized ZSL (GZSL). The common problem with GZSL setting is that the classifier is always biased to wards the seen classes. This reduces the performance for the unseen classes as the unseen examples are often misclassified to one of the seen classes. A simple approach to handle this was proposed in [3], where the authors suggested to reduce the scores for the seen classes. The amount β by which the scores are additively reduced for the seen classes, is a parameter which needs to be tuned. We use the approach of calibrated stacking, and as we are working with distances instead of similarities, we use the modified prediction rule at inference, given by

$$\mathbf{t}^* = \arg\min_{\mathbf{t}, c \in \{S+U\}} \left\{ d_c(\mathbf{x}, \mathbf{t}) + \beta \mathcal{I}(c \in S) \right\}, \quad (8)$$

where, $\mathbf{x}$ can be audio, video or concatenated feature, $\mathcal{I}$ is the indicator function which is 1 when the input condition is true and 0 otherwise. S, U represent the set of seen and unseen classes respectively.

4. Proposed AudioSet ZSL Dataset

A large scale audio dataset, AudioSet [14], was recently released containing segments from in-the-wild YouTube videos (with audio). These videos are weakly annotated with different types of audio events ranging from human and animal sounds to musical instruments and environmental sound. In total, there are 527 audio events, and each video segment is annotated with multiple labels.

The original dataset being highly diverse and rich, is often used in parts to address specific tasks [32, 39].

To study the task of audiovisual ZSL, we construct a subset of the Audioset containing 156416 video segments. We refer to this subset as the `AudioSetZSL`. While the original dataset was multilabel, the example videos were selected such that every video in `AudioSetZSL` has only one label, ie. it is a multiclass dataset. Fig. 3 shows the number of examples for different classes in `AudioSetZSL`, Tab. 1 gives some statistics.

We follow the steps below to create the `AudioSetZSL`: (i) We remove classes with confidence score (for annotation quality) less than 0.7, (ii) we then determine the group of classes that are semantically similar, e.g. animals, vehicles, water bodies. We do so to ensure that the seen and unseen classes for ZSL have some similarities and the task
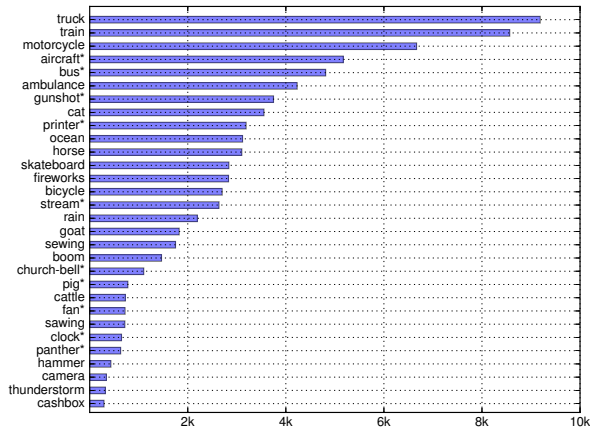
Figure 3. Distribution of the different classes in `AudioSetZSL`. Apart from these three other classes included in the dataset are dog, bird and car containing 12646, 25153 and 38315 examples. The unseen classes are appended with a '*'.

min	max	mean	std. dev.
292	38315	4739.88	7693.10

Table 1. Statistics on the number of examples per class for the `AudioSetZSL` dataset.

is feasible with the dataset. (iii) After selecting the group of classes, we discard correlated classes and classes with very few examples within those groups, obtaining 33 classes. we call two classes correlated if one of them is a subclass of other as per the ontology in original dataset. (iv) We then remove the examples which correspond to more than one of the 33 classes to keep the dataset multiclass. We finally remove the examples that are no longer available on YouTube. Tab. 1 and Fig. 3 give some statistics and more details are in the supplementary document.

To create the *seen, unseen* splits for ZSL tasks, we selected a total of 10 classes spanning all the groups as the zero-shot classes (marked with '*' in Fig. 3). We ensure that the unseen classes have minimal overlap with the Kinetics dataset [40] training classes as we use CNNs pre-trained on that. We do so by not choosing any class whose class embedding similarity is greater than 0.8 with any of the Kinetics train class embeddings in the word2vec space.

We finally split, both the seen and unseen classes, as $60 - 20 - 20$ into train, validation and test sets. We set the protocol to be as follows. Train on the train classes and then test on seen class examples and unseen class examples, both being classified into one of all the classes. The performance measure is mean class accuracies for seen classes and unseen classes and the harmonic mean of these two values, following that in image based ZSL work [10].

5. Experiments

Implementation details The audio network $f_a(\cdot)$ is based on that of [41], and is trained on the spectrogram of the audio clips in the train set of our dataset. We obtain the audio features after seven `conv` layers of the network, and average them to obtain 1024D vector. The video network, denoted as $f_v(\cdot)$ is an inflated 3D CNN network which is pretrained on the Kinetics dataset [40] and a large video dataset of action recognition. We also obtain the video features form the layer before the classification layer and average them to get a vector of 1024D. Finally the text network, denoted as $f_t(\cdot)$ is the well known `word2vec` network pretrained on Wikipedia [42] with output dimension of 300D.

The projection model for text embeddings was fixed to be a single layer network, where as for the audio and video was fixed to be a two layer network, with the output dimensions matching for all. In order to find the seen/unseen class bias parameter β we divide the maximum and minimum possible value of β into 25 equal intervals and then evaluate performances on the *val* set. We chose the best performing β among those.

Evaluation and performance metrics. We report the mean class accuracy (% mAcc) for the classification task and the mean average precision (% mAP) for the retrieval task. The performance for the seen (S) and unseen (U) classes are obtained after classification (retrieval) over all the classes (S and U). The harmonic mean HM of S and U indicates how well the system performs on both seen and unseen categories on average. For classification, we classify each test example, and for retrieval, we perform leave-one-out testing, ie. each test example is considered as a query with the rest being the gallery. The performance reported is (mean class) averaged. A less strict metric (arithmetic mean of S and U) is often used in case of zero-shot retrieval [43, 44], which is not exactly correct.

Methods reported. We report performances of audio and video only methods, ie. only the respective modality is used to test and train. We also report a naive combination by concatenation of features from audio and video modalities before learning the projection to the common space. This method allows zero-shot classification and retrieval only when both the modalities are available, and it does not allow crossmodal retrieval at all. We then report performances with the proposed Coordinated Joint Multimodal Embeddings (CJME) method, when modality attention is used and when it is not used. In either of the cases, we can choose dominant modality (or not) based on the α value (eq. 7). We report with both the cases.

We also compare our approach to two other baseline methods, namely pre-trained features and GCCA. In 'Pretrained' method the raw features obtained from the individual modality pre-trained network are directly used for retrieval as both are of same dimensions. This can be considered as one of the lower bound since no common projection is learned for the different modalities in this case. GCCA
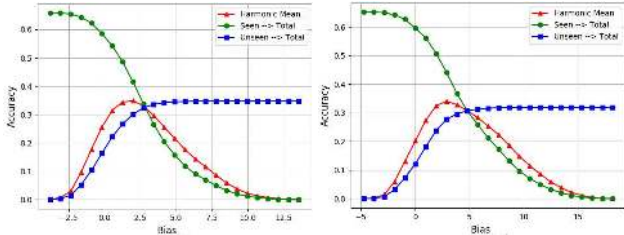
Figure 4. Effect of classification performance for model M1 (left) and model M2 (right) with different values of bias parameter.

Train Modality	Test Modality	S	U	HM
audio	audio	28.35	18.35	22.22
CJME	audio	25.58	20.30	22.64
video	video	43.27	27.11	33.34
CJME	video	41.53	28.76	33.99
both (concat)	both	45.83	27.91	34.70
CJME	both	30.29	31.30	30.79
CJME (no attn)	audio or video	31.72	26.31	28.76
CJME (w/ attn)	audio or video	41.07	29.58	34.39

Table 2. Zero-shot classification performances (% mAcc) achieved with audio only, video only, and both audio and video used for training and test. Note that the audio and video concatenation model requires both the modalities to be available during testing.

[45] or Generalized Canonical Correlation Analysis is the standard extension of the Canonical Correlation Analysis (CCA) method from two-set method to multi-set method, where the correlation between the example pair from each sets are maximized. We use here the GCCA to maximize the correlation between all the three modalities (text, audio and video) for every example triplet in the dataset.

5.1. Quantitative Evaluation

Evaluation of calibrated stacking performance. We have shown the improvement in performance with the approach of calibrated stacking in Fig. 4. This shows the performances with different values of the bias parameter, ie. accuracies for seen and unseen classes, as well as their harmonic means. We observe that the performance increases with the initial increase in bias, and then falls after a certain point as expected. We choose the best performing value of the class bias on the *val* set and then fix it for the experiments on *test* set.

Zero-shot audio-visual classification. Tab. 2 gives the performances of the different models for the task of zero-shot classification. We make multiple observations here. The video modality performs better than the audio modality for the task (33.34 vs. 22.22 HM), which is interesting as the original dataset was constructed for audio event detection. We also observe that when both audio and video modalities are used by simply concatenating the feature from the respective pre-trained networks, the performance increases to 34.70. This shows that adding the audio modality is helpful for zero-shot classification. Our coordinated joint mul-

Model	Test	S	U	HM
pre-trained	T$\rightarrow$ A	3.83	1.66	2.32
GCCA [45]	T$\rightarrow$ A	49.84	2.39	4.56
audio	T$\rightarrow$ A	43.16	3.34	6.20
CJME	T$\rightarrow$ A	48.24	3.32	6.21
pre-trained	T$\rightarrow$ V	3.83	2.53	3.05
GCCA [45]	T$\rightarrow$ V	57.67	3.54	6.67
video	T$\rightarrow$ V	48.62	5.25	9.47
CJME	T$\rightarrow$ V	59.39	5.55	10.15
both (concat)	T$\rightarrow$ AV	63.13	7.80	13.88
CJME	T$\rightarrow$ AV	65.45	5.40	9.97
CJME (no attn)	T$\rightarrow$ A or V	65.74	5.09	9.45
CJME (w/ attn)	T$\rightarrow$ A or V	62.97	5.67	10.41

Table 3. Zero-shot retrieval performances (% mAP) achieved by models when audio only, video only, and both audio and video modalities are used for training and test. Note that the audio and video concatenation based model requires both modalities at test time also and can not predict using any single one.

timodal embeddings (denoted CJME in the table) improves the performance of video and audio only models on the respective test sets by modest but consistent margins. This highlights the efficacy of the proposed method to learn joint embeddings which are comparable (slightly better) than individually trained models.

The performance of the proposed method is lower without attention learning and selective modality based test time prediction cf. the concatenated input model (30.79 vs. 34.70), but is comparable to it when trained and tested with attention (34.39). Also, when we do not train for attention but use selective modality based prediction the performance falls (28.76). Both these comparisons validate that the modality attention learning is an important addition to the base multimodal embedding learning framework.

Zero-shot audio-visual retrieval. Tab. 3 compares the performances of different models for the task of zero-shot retrieval. The performance on the unseen classes are quite poor, albeit it is approximately three times the baseline pre-trained performance. This is because of the bias towards the seen classes in generalized ZSL. This happens for classification setting as well but is corrected for explicitly by reducing the scores of the seen classes. However, in a retrieval scenario, since the class of the gallery set member is not known in general, such correction can not be applied. We tried classifying the gallery sets first and then applying the seen/unseen class bias correction, however that did not improve results possibly because of erroneous classifications.

We observe, from Tab. 3, similar trends as with zero-shot classification. The proposed CJME performs similar to audio only (6.20 vs. 6.21) and slightly better than video only (9.47 vs. 10.15) models but consistently outperforms both pre-trained and GCCA model. Compared to the audio and video features concatenated model, the performance without modality attention based training are lower (13.88

Model	Test	S	U	HM
pre-trained	audio → video	3.61	2.37	2.86
GCCA [45]	audio → video	22.12	3.65	6.26
CJME	audio → video	26.87	4.31	7.43
pre-trained	video → audio	4.22	2.57	3.19
GCCA [45]	video → audio	26.68	2.98	5.26
CJME	video → audio	29.33	4.35	7.58

Table 4. Zero-shot crossmodal retrieval performances (% mAP).

$\mathcal{L}_{AV}$	$\mathcal{L}_{TA}$	$\mathcal{L}_{TV}$	S	U	HM
✗	✗	✓	1.26	10.13	2.24
✗	✓	✗	3.00	4.18	3.49
✗	✓	✓	31.20	28.47	29.77
✓	✗	✓	30.39	27.31	28.76
✓	✓	✗	30.07	25.06	27.33
✓	✓	✓	33.29	28.18	30.53

Table 5. Ablation study to verify the contribution of different loss terms on val set. Performances for proposed CJME (with attention) method on zero-shot classification (% mAcc)

vs. 9.47) which improve upon using attention at training (10.41), albeit staying a little lower cf. similar in the classification case. We thus conclude that CJME is as good as audio only or video only model and is competitive cf. concatenated features model, while allowing crossmodal retrieval, which we evaluate next.

Crossmodal retrieval. Since CJME learns to embed both audio and video modality in a common space, it allows for doing crossmodal retrieval from audio to video and vice-versa. Tab. 4 gives the performances of such crossmodal retrieval from audio and video domains. We observe that the retrieval accuracy in the case of crossmodal retrieval are 7.43 and 7.58 for audio to video and video to audio respectively. Due to the inability to do seen/unseen class bias correction, we observe a large gap between the retrieval performance of seen classes cf. unseen classes, which stays true in the case of crossmodal retrival as well. The performance is still three times better than the raw pre-trained features. We believe these are encouraging initial results on the challenging task of audio-visual crossmodal retrieval on real world unconstrained videos in zero-shot setting.

Ablation of the different loss components. Tab. 5 gives the performances on val set in the different cases when we selectively turn off different combination of losses in the optimization objective eq. 4. We observe that all three losses contribute positively towards the performance. When either of the triplet loss is turned off, the performance drastically fall to ~ 3, but when the crossmodal audio-video loss is added with one of the triplet losses turned off, they recover to reasonable values ~ 28. Compared to the final performance of 30.53, when the text-audio, text-video and audio-video losses are turned off, the performances fall to 28.76, 27.33 and 29.77 respectively. Thus we conclude that each component in the loss function is useful and that the networks (which are already pre-trained on auxiliary classi-

fication tasks) need to be trained for the current task to give meaningful results.

5.2. Comparison with state of the art methods

We address both possible issues, ie. (i) our implementation is competitive w.r.t. other methods on standard datasets, and (ii) how do other methods compare on the proposed `AudioSetZSL` dataset, by providing additional results. Tab. 6 gives the performance of our implementation on other datasets (existing method performances are taken from Xian et al. [10]).

We observe that our method is competitive to other methods on an average. Tab. 7 gives the classification performance of other methods using our features on the proposed dataset. We see that our method performs better than many existing methods (eg. ALE 33.0 vs. CJME 34.4). Hence we conclude that our implementation and method, both, perform comparable to existing appearance based ZSL methods. Tab. 7 also shows that adding audio improves the video only ZSL from 33.3 to 34.4 HM.

In these comparisons, we have not included some of the recent generative approaches [2, 9] which handles the task by conditional generation of examples form the unseen classes. Although these approaches increase the performance but they come with the drawback of soft-max classification, which requires the classifier to be trained form scratch once again if a new class is added to the existing setup at test time. This also requires saving all the training data for generative approaches while in the projection based methods, this is not required.

5.3. Qualitative evaluation

Fig. 5 shows qualitative crossmodal retrieval results for all three pairs of modalities, ie. text to audio/video, audio to video and video to audio. We see that method makes acceptable mistakes, eg. for the car text query one of the audio retrieval contains motorbike due to the similar sound, for the bird video query the wrong retrieval is a cat purring sound which is similar to a pigeon sound. In the unseen class case, the bus text query return car, train and truck audio as the top false positives. Easier and distinct cases such as gunshot audio query gives very good video retrievals. We encourage the readers to look at the result videos available at `https://www.cse.iitk.ac.in/users/kranti/avzsl.html` for a better understanding of qualitative results.

6. Conclusion

We presented a novel method, which we call Coordinated Crossmodal Joint Embeddings (CJME), for the task of audio visual zero-shot classification and retrieval of videos. The method learns to embeds audio, video and text into a

Method	SUN			CUB			AWA1			AWA2		
	U	S	HM	U	S	HM	U	S	HM	U	S	HM
CONSE [7]	6.8	39.9	11.6	1.6	72.2	3.1	0.4	88.6	0.8	0.5	90.6	1.0
DEVISE [4]	16.9	27.4	20.9	23.8	53.0	32.8	13.4	68.7	22.4	17.1	74.7	27.8
SAE [17]	8.8	18.0	11.8	7.8	54.0	13.6	1.8	77.1	3.5	1.1	82.2	2.2
ESZSL [5]	11.0	27.9	15.8	12.6	63.8	21.0	6.6	75.6	12.1	5.9	77.8	11.0
ALE [8]	21.8	33.1	26.3	23.7	62.8	34.4	16.8	76.1	27.5	14.0	81.8	23.9
CJME	30.2	23.7	26.6	35.6	26.1	30.1	29.8	47.9	36.7	51.9	36.8	43.1

Table 6. Comparison with existing methods on standard datasets (projection based methods only, see Sec. 5.2 for details)

Method	Modality		S	U	HM
	train	test			
CONSE [7]	video	video	48.5	19.6	27.9
DEVISE [4]	video	video	39.8	26.0	31.5
SAE [17]	video	video	29.3	19.3	23.2
ESZSL [5]	video	video	33.8	19.0	24.3
ALE [8]	video	video	47.9	25.2	33.0
CJME	video	video	43.2	27.1	33.3
CJME	both	video	41.5	28.8	33.9
CJME	both	both	41.0	29.5	34.4

Table 7. Comparison with existing methods on proposed dataset (projection based methods only, see Sec. 5.2 for details)
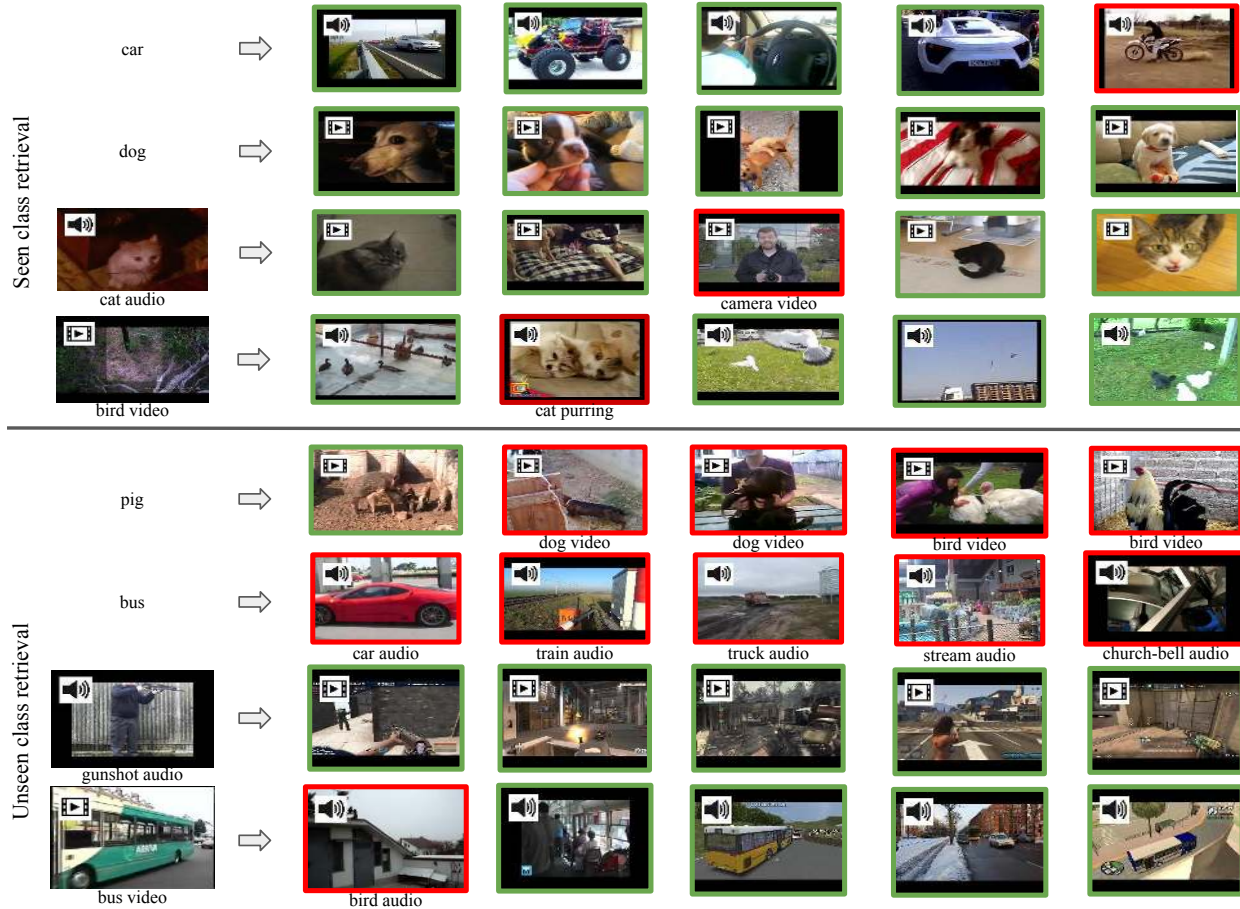

Figure 5. Qualitative crossmodal retrieval results with the proposed method. Each block of two rows from top to bottom corresponds to text to audio, text to video, audio to video and video to video respectively. The small icons on the left top of each image indicates the modality considered for that specific video. Please see detailed results video in the supplementary material.

common embedding space and then performs nearest neighbor retrieval in that space for classification and retrieval. The loss function we propose has three components, two bimodal text-audio and text-video triplet losses, and an audio-video crossmodal similarity based loss. Motivated by the fact that the two modalities might carry different amount of information for different examples, we also proposed a modality attention learning framework. The attention part learns to predict the dominant modality for the task, ie. if the object is occluded but the audio is clear, and base the prediction on that modality only. We reported extensive experiments to validate the method and showed advantages of the method over baselines, as well as demonstrated crossmodal retrieval which is not possible with the baseline methods. We also constructed a dataset appropriate for the task which is a subset of a large scale unconstrained dataset for audio event detection in video.

References

[1] Zeynep Akata, Scott Reed, Daniel Walter, Honglak Lee, and Bernt Schiele. Evaluation of output embeddings for fine-grained image classification. In *Computer Vision and Pattern Recognition*, pages 2927–2936, 2015.

[2] Yongqin Xian, Tobias Lorenz, Bernt Schiele, and Zeynep Akata. Feature generating networks for zero-shot learning. In *Computer Vision and Pattern Recognition*, 2018.

[3] Wei-Lun Chao, Soravit Changpinyo, Boqing Gong, and Fei Sha. An empirical study and analysis of generalized zero-shot learning for object recognition in the wild. In *European Conference on Computer Vision*, pages 52–68. Springer, 2016.

[4] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems*, pages 2121–2129, 2013.

[5] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In *International on Machine Learning*, pages 2152–2161, 2015.

[6] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2014.

[7] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013.

[8] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *Pattern Analysis and Machine Intelligence*, 38(7):1425–1438, 2016.

[9] V Kumar Verma, Gundeep Arora, Ashish Mishra, and Piyush Rai. Generalized zero-shot learning via synthesized examples. In *Computer Vision and Pattern Recognition*, pages 4281–4289, 2018.

[10] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *Transactions on Pattern analysis and Machine Intelligence*, 2018.

[11] Ashish Mishra, Vinay Kumar Verma, M Shiva Krishna Reddy, S Arulkumar, Piyush Rai, and Anurag Mittal. A generative approach to zero-shot and few-shot action recognition. In *Winter on Applications of Computer Vision*, pages 372–380. IEEE, 2018.

[12] Jie Qin, Li Liu, Ling Shao, Fumin Shen, Bingbing Ni, Jiaxin Chen, and Yunhong Wang. Zero-shot action recognition with error-correcting output codes. In *Computer Vision and Pattern Recognition*, pages 2833–2842, 2017.

[13] Xun Xu, Timothy M Hospedales, and Shaogang Gong. Multi-task zero-shot action recognition with prioritised data augmentation. In *European on Computer Vision*, pages 343–359. Springer, 2016.

[14] 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 *International on Acoustics, Speech and Signal Processing*, pages 776–780. IEEE, 2017.

[15] Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems*, pages 935–943, 2013.

[16] Yongqin Xian, Zeynep Akata, Gaurav Sharma, Quynh Nguyen, Matthias Hein, and Bernt Schiele. Latent embeddings for zero-shot classification. In *Computer Vision and Pattern Recognition*, pages 69–77, 2016.

[17] Elyor Kodirov, Tao Xiang, and Shaogang Gong. Semantic autoencoder for zero-shot learning. In *Computer Vision and Pattern Recognition*, pages 3174–3183, 2017.

[18] Debasmit Das and CS Lee. Zero-shot image recognition using relational matching, adaptation and calibration. *arXiv preprint arXiv:1903.11701*, 2019.

[19] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. *arXiv preprint arXiv:1804.04340*, 2018.

[20] Pengkai Zhu, Hanxiao Wang, Tolga Bolukbasi, and Venkatesh Saligrama. Zero-shot detection. *arXiv preprint arXiv:1803.07113*, 2018.

[21] Shafin Rahman, Salman Khan, and Fatih Porikli. Zero-shot object detection: Learning to simultaneously recognize and localize novel concepts. *arXiv preprint arXiv:1803.06049*, 2018.

[22] Vinay Kumar Verma and Piyush Rai. A simple exponential family framework for zero-shot learning. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 792–808. Springer, 2017.

[23] Wenlin Wang, Yunchen Pu, Vinay Kumar Verma, Kai Fan, Yizhe Zhang, Changyou Chen, Piyush Rai, and Lawrence Carin. Zero-shot learning via class-conditioned deep generative models. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[24] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[25] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Lo-geswaran, Bernt Schiele, and Honglak Lee. Genera-tive adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[26] Xinjian Li, Siddharth Dalmia, David R. Mortensen, Florian Metze, and Alan W Black. Zero-shot learning for speech recognition with universal phonetic model, 2019.

[27] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Sound-net: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems*, pages 892–900, 2016.

[28] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European on Computer Vision*, pages 801–816. Springer, 2016.

[29] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *International on Computer Vision*, pages 609–617. IEEE, 2017.

[30] Andrew Owens and Alexei A Efros. Audio-visual scene analysis with self-supervised multisensory features. *arXiv preprint arXiv:1804.03641*, 2018.

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

[32] Ruohan Gao, Rogerio Feris, and Kristen Grauman. Learning to separate object sounds by watching unlabeled video. *arXiv preprint arXiv:1804.01665*, 2018.

[33] Relja Arandjelović and Andrew Zisserman. Objects that sound. *arXiv preprint arXiv:1712.06651*, 2017.

[34] Sanjeel Parekh, Alexey Ozerov, Slim Essid, Ngoc Duong, Patrick Pérez, and Gaël Richard. Identify, locate and separate: Audio-visual object extraction in large video collections using weak supervision. *arXiv preprint arXiv:1811.04000*, 2018.

[35] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Computer Vision and Pattern Recognition*, pages 4358–4366, 2018.

[36] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. Seeing voices and hearing faces: Cross-modal biometric matching. In *on Computer Vision and Pattern Recognition*, pages 8427–8436, 2018.

[37] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech sepa-ration. *ACM Transactions on Graphics*, 37(4):112, 2018.

[38] Einat Kidron, Yoav Y Schechner, and Michael Elad. Pixels that sound. In *Computer Vision and Pattern Recognition*, volume 1, pages 88–95. IEEE, 2005.

[39] Emmanuel Vincent INRIA Benjamin Elizalde, Bhiksha Raj. Large-scale weakly supervised sound event detection for smart cars, 2017.

[40] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Com-puter Vision and Pattern Recognition*, 2017.

[41] Anurag Kumar, Maksim Khadkevich, and Christian Fügen. Knowledge transfer from weakly labeled audio using con-volutional neural network for sound events and scenes. In *International on Acoustics, Speech and Signal Processing*, pages 326–330. IEEE, 2018.

[42] Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training dis-tributed word representations. In *International on Language Resources and Evaluation*, 2018.

[43] Sasi Kiran Yelamarthi, Shiva Krishna Reddy, Ashish Mishra, and Anurag Mittal. A zero-shot framework for sketch based image retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 300–317, 2018.

[44] Anjan Dutta and Zeynep Akata. Semantically tied paired cycle consistency for zero-shot sketch-based image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5089–5098, 2019.

[45] Jon R Kettenring. Canonical analysis of several sets of vari-ables. *Biometrika*, 58(3):433–451, 1971.