博士論文（要約）

Generating Multimodal Explanations of Visual Categorization

based on Mutual Information Maximization

（相互情報量最大化に基づく複数モダリティによる

視覚情報の分類に対する説明）

兼平 篤志

Abstract

This study addresses generating multimodal explanations for the categorization of visual information. Although most researches focus on explanation by the single modal information, it can be insufficient when what is explained cannot be represented by single modal, or when interpretability degrades when attempting to output faithful explanation. Even though an existing work treats multimodality on generating explanations, it requires supervised information of correct explanations for all modalities. Collecting such dataset is costly, and even impossible for some modalities, preventing the application of the explanation system to the real problem. An important aspect of multimodal explanations is complementarity, that is, how one explanation improve the quality of the others. Thus, as well as general requirements for generated explanations: (a) interpretable, (b) fidelity to the target, the system generating multimodal explanations are required to be (c) applicable to modals regardless of with/without supervised information, and (d) able to generate explanations where different modals are complemental to each other. To satisfy (a)-(d) simultaneously, we propose a novel framework based on information theory. Defining distributions of variables to explain and to be explained, explanations holding high interaction information is selected. To apply it to different types of explanations utilizing different domains, we propose practical algorithms corresponding to each case of realistic application, and demonstrate their effectiveness by conducted experiments.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

1.1 Background

The explanation is an intellectual act of replacing the basis of a certain fact or decision with a different expression interpretable for the other. Two subjects involved in the act: who explains and who is explained, and we ask for an explanation when (1) we need to obtain knowledge on the event, or when (2) we require to check whether the other subject makes a conclusion by the correct reasoning.

Explanations are also important for visual recognition. Considering the recognition process of by humans or machines where a visual instance, e.g., image or video, is classified into a certain category, the process is often opaque. Those who obtain the recognition result requires an explanation mainly in the following two situations related to the cases mentioned above:

- when the decision made by one subject is not obvious by the other, and he/she demands to acquire detailed knowledge of the event. A possible situation is that the decision is made by the expert, such as when a medical doctor assigns labels of diagnosis to medical images. Another situation is when the decision is subjective when asking someone's preference or sentiment to images.

- when one would like to verify whether the result is obtained by the correct reasoning. For example, the decision process is complicated and hard to interpret by humans in the object recognition system based on deep Convolutional Neural Networks (CNNs). Even though it performs well in the test dataset, it does not necessarily mean the system works desirably in the real application because the possibility of exploiting dataset-bias remains. In such situation, explanation helps us to verify trustability of the reasoning process.

Existing researches have focused on generating explanations on visual data. Some of them attempt to provide explanations for complicated machine learning model, by exploiting the natural language or parts of target instance. On the other hand, researches for extracting representative and discriminative information from the visual instances labeled by humans can also be regarded as a kind of explanation.

Almost all researches utilize the information on the single modal; however, it can be insufficient in some situations. A possible situation is when what is explained cannot be represented by single modal, or when interpretability degrades when attempting to output faithful explanation. For example, it may require tons of words for explaining the reason for an image being categorized to a specific class utilizing only natural language.

This study addresses exploiting multimodal information for the explanations to overcome the difficulty. Even though an existing work [56] treats multimodality on generating explanations, it requires supervised information of correct explanations for all modalities. Collecting such dataset is costly, and even impossible for some modalities, preventing the application of the explanation system to the real problem. Moreover, it only treats the positiveness of the label, and cannot deal with negativeness of it, which is another important application of the explanation.

One important factor on the multimodal explanation is complementarity. By combining multiple information holding different characteristics, it can be expected that they complement to each other. In the example stated above, it may improve the fidelity and/or interpretability of the explanation not only by adhering to the language but also by utilizing visual information such as pointing appropriate regions in the target image. This is a simple example of achieving complementarity of visual and linguistic information, provided that they are good at representing low-level and high-level information respectively.

In summary, we require a framework for multimodal explanations, which can deal with several modalities regardless with/without supervised information, and which can achieve complementarity of different modalities.

1.2 Scope of Thesis

In this thesis, we especially focus on the classification problem as the target of the explanation, where a visual instance $\mathbf{x}$ (e.g., image or video) is categorized to a class $\mathbf{y}$ by the process $\mathbf{f} : \mathbf{x} \rightarrow \mathbf{y}$. Under this premise, the task of explanation is divided by the two axes as follows:

- The difference of the decision process $\mathbf{f}$. One is when humans perceive visual information and categorize it. The other is by machines. We especially deal with the

complicated machine learning models such as deep CNNs utilized as the de-facto standard in the current visual recognition task.

- The difference of dealing with positiveness/negativeness of the label **y**. In other words, the difference of explaining "why **x** is categorized to **y**" or "why **x** is **not** categorized to **y**."

By these axes, explanation task is divided into four parts, those are, machine-positiveness / machine-negativeness / human-positiveness / human-negativeness. This thesis tackles former three i.e., machine-positiveness / machine-negativeness / human-positiveness, provided that the remaining one is considered to be achieved by combining them.

1.3 Objective

The goal of this thesis is to build a system that can generate explanations with multimodal information for categorization of a visual instance, and we propose a general framework to achieve it.

We first need to discuss the requirements of the system. The desired explanation is the one which satisfies following two properties as:

(a) explanations should be interpretable for humans, and

(b) explanations should have fidelity to the target to be explained. As we especially treat classification problem represented by $\mathbf{x} \rightarrow \mathbf{y}$ in this study, the fidelity can be rephrased by discriminativeness. In other words, the explanation should retain information useful to identify the category to which the instance is predicted to belong.

In addition to generating explanations satisfying above, requirements of the proposed framework for generating multimodal explanations are:

(c) it should be applicable to modals regardless with/without supervised information,

(d) its outputted explanations of different modals should be complemental to each other.

We propose a novel framework for generating multimodal explanations based on the information theory, especially on the maximization of the interaction information, which is an extension of the mutual information defined on more than two variables.

To guarantee the output explanations are (a) interpretable for humans, we limit modalities used for explanations. As the interpretability depends heavily on the humans' ability of perception, this work makes an assumption that a set of (or parts of) linguistics, real images,

and videos are interpretable. To achieve (b) discriminativeness of output explanations to the category information, we generate it by taking the interaction information into account. We define distributions of variables to explain, and to be explained. The latter represents the process of a instance being categorized by the subject, e.g., humans or machines. The interactive information defined on these distributions provides the quantity measuring how the category is identifiable from explanations. By selecting explanations holding high interaction information, discriminative explanations can be generated. Further, considering (c), because the explanations are generated to make the category information identifiable, not to attempt to match the supervised information in our framework, it is applicable to modals without supervised information. The (d) complementarity is naturally satisfied by the definition of interaction information. The interaction information is defined recursively by the difference between interactive information with/without being conditioned on one variable, providing a natural definition on the complementarity: the increase of the dependency of the other explanations to the target category when one explanation is conditioned. As observed, this extension enables to satisfy (a)-(d) simultaneously. Details are discussed in the subsequent chapter.

We apply the proposed framework to each case in the realistic applications of explanation task mentioned above as:

- whether the decision is made by humans, or made by machines, and
- whether the explanation is made for the positiveness, or the negativeness of the target category,

and propose algorithms specific to them. We also demonstrate the effectiveness of these algorithms by the experiments. Contributions of this thesis are summarized as follows:

- propose a general framework for generating multimodal explanations based on the information theory,
- propose novel algorithms based on the proposed framework for applying it to different kinds of realistic explanation tasks, and
- demonstrate the effectiveness of proposed algorithms by the experiments.

1.4 Structure of Thesis

In this thesis, we propose a framework for generating multimodal explanations on a prediction of visual information to the category. The remainder of this paper is organized as follows. In

Chapter 2, existing works related to our task are discussed. In Chapter 3, after illustrating the proposed framework for explanation generations, we describe two axes by which the explanation task is divided: (1) whether the decision is made by humans, or made by machines, and (2) whether the explanation is made for the positiveness or the negativeness of the target category. From Chaper 4 to Chapter 6, we propose practical algorithms when applying the framework to realistic applications following that taxonomy on the explanation task. Particularly, we focus on the generating explanations for

- the positiveness of labels predicted by machines in Chaper 4,
- the negativeness of labels predicted by machines in Chaper 5,
- the positiveness of labels predicted by humans in Chaper 6,

and we demonstrate the effectiveness of the proposed method in each task by the conducted experiments. Finally, we conclude our work and discuss future works in Chapter 7.

Chapter 2

Related Work

Recently, several works have addressed generating explanations for the prediction of visual information. The main motivation of them are either or / both of (1) obtaining detailed knowledge, (2) validating trustability of the reasoning. In this chapter, we discuss previous works related to our task.

2.1 Single modal explanations

Almost all existing researches focus on the explanation utilizing the single modal information. We divide them by two axes mentioned in the Section 1.1 as:

- whether the decision is made by humans, or made by machines, and
- whether the explanation is made for the positiveness, or the negativeness of the target category.

2.1.1 Explanation for the positiveness of the machines' prediction

The visual cognitive ability of machines has improved significantly primarily because of the recent development in deep-learning techniques. Owing to its high complexity, the decision process is inherently a black-box; therefore, many researchers have attempted to make a machine explain the reason for the decision to verify its trustability.

The primary stream is visualizing where the classifier weighs for its prediction by assigning an importance to each element in the input space by rule-based [66, 5, 83, 63, 85, 19, 86] approach or learning-based approach [8, 11]. Methods in the former category decide the rule of propagating importance between two layers, and apply it recursively from the output prediction to the input. For example, [63] proposed to calculate the gradient of the

output with regard to the elements of mid-level features in interest by back-propagation, and to aggregate them following the multiplication to the input to make the importance. [5] proposed a technique Layer-wise Relevance Propagation (LRP) which propagates the element of the output of each layer to the input by the propagation rule determined by the kind of input domain. Also, recently [86] attempted to enhance interpretability of the result by decomposing the importance propagated between layers to those belong to human-interpretable concepts, by train the mapping function on the auxiliary densely annotated dataset.

Methods belong to the latter attempts to learns the instance-wise importance of elements with an auxiliary model, and they are further divided by how the model is trained, those are element-wise or instance-wise. [60] is the representative research belonging to the first category. Under an assumption that the decision function around the target input can be approximated as locally linear, and they approximated the complex decision function by training the sparse linear regression model by pairs of input data dropping some elements and its prediction. The assumption that complex decision function can be considered as locally linear even though it cannot be approximated globally by the simple function is often utilized in other works such as gradient-based method mentioned above. Although the above approaches aim to show the region where the classifier weights in the target, the goal of [38] was to detect a training sample on which the prediction heavily relies on. They proposed an approach to use the influence function to assess the influence of the sample's absence on the classifier's prediction.

Unlike the former approaches, the latter approaches train instance-wise model which predicts the importance of the input element. A merit of learning instance-wise importance is one can obtain the importance relative to other samples even though the former focus on the importance of element only inside the target sample. [8] proposed to utilize the mutual information. In their work, the method holds two different neural network models which predict the importance of each element of input as well as the classifier which predicts the class the masked input belongs to. [11] demonstrated that the importance predicted in a similar way can be used well as saliency map.

The main goal of these works is to show where the model actually "looks" in a human-interpretable manner, it is important to show the region where the model focuses for prediction rather than whether the prediction is true for humans. The evaluation is often performed by the investigating before/after the output of the model when the element considered to be important is changed.

As a different stream, some works trained the generative model that projects the mid-level feature of the classifier to be explained onto the other information obtained from outside in a

post-hoc manner. The representative work is [29], which exploits natural language. They claimed that the explanation by natural language is different from the description in that the explanation is required to be both of discriminative to class and relevant to the input target, although the description only needs to satisfy relevance. To guarantee the explanation is discriminative to the target category, they proposed to use an additional classification model as well as explanation generation model, which verifies whether the generated explanation is class specific. Owing to the inability of backpropagation, they are trained by reinforcement learning in their work. Also, the work utilizing multimodal [56] information belong to this category. It exploits not only natural language but also visual information, that is a pointing supervision on the image. As described later section, as it requires supervised information of the correct explanation for all modalities, collecting data is costly or impossible for some domain. True explanations, as by humans, are expected to be generated regardless of the type of model in these type of explanation. In this sense, [61] can be considered as a part of this stream. It introduced a loss which imposes a penalty on the difference from the ground-truth explanation in addition to the ordinal classification loss when training classifier to render it not only predict correctly but also predicts with *right reason*. In most case, as the desired output is obtained, the quantitative evaluation is often performed by comparison with ground-truth supervision.

2.1.2 Explanation for the negativeness of the machines' prediction

Few works attempted to generate explanation being aware to the class different from the one the classifier provides. [3] stated an application for grounding visual explanation to counterfactual explanation, that is the explanation of "why the target is not predicted to a specific category", where the textual explanations are generated by comparing the output of generated explanations for target sample and the nearest sample to it. [73] proposed the method to compute the minimum change of the input leading the different decision of classifier (e.g., positive $\rightarrow$ negative), and applied it to the improvement of Internet advertisement. They mainly focused on Random Forest models whose input element is interpretable for a human. Thus, this can not be easily computer vision task where the input is usually raw pixels or high-dimensional feature, whose each element does not have a meaning itself. [62] resorted on Generative Adversarial Networks (GANs), which is known to perform well in the image generation task, to change the input image in a meaningful way, by which the target classifier's decision changes. It can be regarded as a kind of explanation for the negative class.

2.1.3 Explanation for the positiveness of the humans' prediction

Prototype selection [68, 43, 31, 13–15] can be considered as example-based explanations. The essential idea of these researches it to extract information that is representative to the sample distribution, and is discriminative to the category the target image belongs. For example, [68] proposed an algorithm extracting the center of clusters constructed by discriminative clustering [80], and [31] applied this idea to the video. Subsequently, [13] proposed a more efficient algorithm using mode seeking. [43] focused on the CNN features and extended the association rule mining, which is a pattern mining algorithm aiming to discover a set of if-then rules. In other words, they attempt to obtain examples that represent $p(\mathbf{x}|c)$, which is the distribution of sample $\mathbf{x}$ conditioned on the category c. Our work is different in that we attempt to explain the black-box posterior distribution $p(c|\mathbf{x})$ such as that represented by deep CNN.

Similarly, [48], which tackled the visual explanation task from the viewpoint of machine teaching, is also a type of example-based explanations. In the machine teaching, regarding weak classifier as humans model, the teacher model, that knows all answers on given dataset, learns to teach how to the student efficiently improve the classification performance. In addition, providing a feedback of category label to students as in the ordinal machine teaching setting, they provide the pixel-level importance on the image to enhance interpretability, and they state the application to visual explanation tasks.

Although most studies are focused on the single modality; however, it can be insufficient in some situations. A possible situation is when what is explained cannot be represented sufficiently by single modal, or when interpretability degrades when attempting to output faithful explanation. To overcome this difficulty, our work particularly treats multimodal information for the explanation.

2.2 Multimodal explanations

There exist a research which has treated multimodality for explanation [56], that is visual and linguistic. It requires supervised information for all modalities of explanations. it requires supervised information of correct explanations for all modalities. Collecting such dataset is costly, and even impossible for some modalities, preventing the application of the explanation system to the real problem. Moreover, it only treats the positiveness of the label, and cannot deal with negativeness of it, which is another important application of the explanation.

Chapter 3

Methodology

3.1 Framework for generating multimodal explanations

The goal of this thesis is to build a system that can generate explanations with multimodal information for categorization of a visual instance. Requirements of our explanation system and its output satisfy are as follows:

(a) explanations should be interpretable for humans, and

(b) explanations should have fidelity to the target to be explained. As we especially treat classification problem represented by $\mathbf{x} \rightarrow \mathbf{y}$ in this study, the fidelity can be rephrased by discriminativeness. In other words, the explanation should retain information useful to identify the category to which the instance is predicted to belong.

(c) it should be applicable to modals regardless with/without supervised information,

(d) its outputted explanations of different modals should be complemental to each other.

Denoting the variable which is the target of the explanation as $\mathbf{y}$, and the variables representing different modality utilized for the explanation as $\mathbf{e}_1$, $\mathbf{e}_2$,...,$\mathbf{e}_N$. We consider the distributions of these variables, and attempt to maximize the interaction information, which is an extension of mutual information defined on more than two variables as:

$$\max_{\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_N} \mathrm{MI}(\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{e}_N,\mathbf{y}|\mathbf{x}) \tag{3.1}$$

where $\mathrm{MI}(\mathbf{e}_1,\mathbf{e}_2,\ldots,\mathbf{y}|\mathbf{x})$ is conditional interactive information. This framework can satisfy abovelisted requirements simultaneously.

To guarantee the output explanations are (a) interpretable for humans, we limit modalities used for explanations. As the interpretability depends heavily on the humans' ability of

perception, this work makes an assumption that a set of (or parts of) linguistics, real images, and videos are interpretable. To achieve (b) discriminativeness of output explanations to the category information, we generate it by considering interaction information. We define distributions of variables to explain $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N$, and to be explained $\mathbf{y}$. The latter represents the process of an instance being categorized by the subject, e.g., humans or machines. The interactive information defined on these distributions provides the quantity measuring how the category is identifiable from explanations. By selecting the explanation holding high interaction information, the discriminative explanation can be generated. To observe it, we consider utilizing single modal, where a special case of $N = 1$. The mutual information of two variables can be written as:

$$\mathrm{MI}(\mathbf{e}, \mathbf{y}|\mathbf{x}) = \mathbb{E}_{\mathbf{x},\mathbf{y},\mathbf{e}}\left[\frac{p(\mathbf{e}|\mathbf{x},\mathbf{y})}{p(\mathbf{e}|\mathbf{x})}\right] = \mathbb{E}_{\mathbf{x},\mathbf{y},\mathbf{e}}\left[\frac{p(\mathbf{e}|\mathbf{x},\mathbf{y})}{\mathbb{E}_{\mathbf{y}}\left[p(\mathbf{e}|\mathbf{x},\mathbf{y})\right]}\right]. \tag{3.2}$$

Intuitively, explanations $\mathbf{e}$, by which the denominator gets small and the numerator gets large, is discriminative information, which is specific for the target category and not important for other possible categories, is selected.

In this study, we further considering because the explanations are generated to make the category information identifiable, not to attempt to match the supervised information in our framework, it is applicable to modals without supervised information.

$$\mathrm{MI}(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_N, \mathbf{y}|\mathbf{x}) = \mathrm{MI}(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{N-1}, \mathbf{y}|\mathbf{x}, \mathbf{e}_N) - \mathrm{MI}(\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{N-1}, \mathbf{y}|\mathbf{x}) \tag{3.3}$$

The interaction information is defined recursively by the difference between interactive information with/without being conditioned on one variable. For other variables, it can be written in the same way.

interaction information provides a natural definition on complementarity: the increase of the dependency between $\mathbf{y}$ and explanations $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{N-1}$ of the other explanations to the target category when one explanation $\mathbf{e}_n$ is conditioned.

Chapter 6

Multimodal Explanations for the Positiveness of Humans' Prediction

This paper introduces a novel variant of video summarization, namely building a summary that depends on the particular aspect of a video the viewer focuses on. We refer to this as *viewpoint*. To infer what the desired viewpoint may be, we assume that several other videos are available, especially groups of videos, e.g., as folders on a person's phone or laptop. The semantic similarity between videos in a group vs. the dissimilarity between groups is used to produce viewpoint-specific summaries. For considering *similarity* as well as avoiding redundancy, output summary should be (A) diverse, (B) representative of videos in the same group, and (C) discriminative against videos in the different groups. To satisfy these requirements (A)-(C) simultaneously, we proposed a novel video summarization method from multiple groups of videos. Inspired by Fisher's discriminant criteria, it selects summary by optimizing the combination of three terms (a) inner-summary, (b) inner-group, and (c) between-group variances defined on the feature representation of summary, which can simply represent (A)-(C). Moreover, we developed a novel dataset to investigate how well the generated summary reflects the underlying *viewpoint*. Quantitative and qualitative experiments conducted on the dataset demonstrate the effectiveness of proposed method.

6.1 Introduction

Owing to the recent spread of Internet services and inexpensive cameras, an enormous number of videos have become available, making it difficult to verify all content. Thus, video summarization, which compresses a video by extracting the *important* parts while avoiding redundancy, has attracted the attention of many researchers.

Figure 6.1: Many types of summaries can exist for one video based on the *viewpoint* toward it.

The information deemed *important* can be varied based on the particular aspect the viewer focuses on, which hereafter we will refer to as *viewpoint* in this paper[1]. For instance, given the video in which the running events take place in Venice, as shown in Fig. 6.1, if we watch it focusing on the "kind of activity," the scene in which many runners come across in front of the camera is considered to be important. Alternatively, if the attention is focused on "place," the scene that shows a beautiful building may be more important. Such *viewpoints* may not be limited to explicit ones stated in the above examples, and in this sense, the optimal summary is not necessarily determined in only one way.

Most existing summarization methods, however, assume there is only one optimal for one video. Even though the variance between subjects are considered by comparing multiple human-created summaries during evaluation, it is difficult to determine how well the *viewpoint* is considered.

Although several different ways may exist for interpreting a *viewpoint*, this paper takes the approach of dealing with it by considering the *similarity*, which represents what we feel is similar or dissimilar, and has a close relationship with the *viewpoint*. For example, as shown in Fig. 6.2, "running in Paris" is closer to "running in Venice" than "shopping in Venice" from the *viewpoint* of the "kind of activity," but such a relationship will be reversed when the *viewpoint* changes to "place." Here, we use the word *similarity* to indicate the one that captures semantic information rather than the appearance, and importantly, it is changeable

[1]Note it does not mean the physical position.

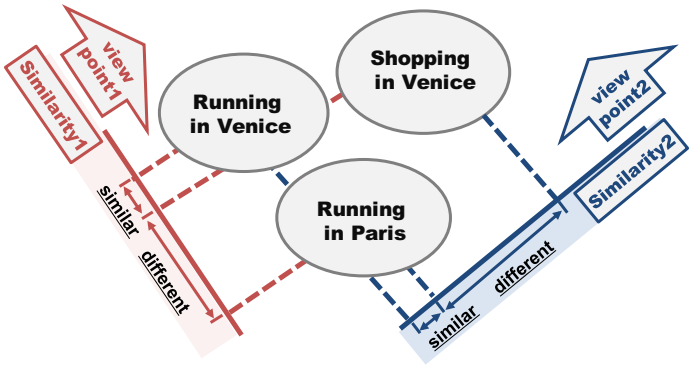


Figure 6.2: Conceptual relationship between a *viewpoint* and *similarity*. This paper assumes a *similarity* is derived from a corresponding *viewpoint*.

depending on the *viewpoint*. We aim to generate a summary considering such *similarities*. A natural question here is "where does the *similarity* come from?"

We may be able to obtain it by asking someone whether two frames are similar or dissimilar for all pairs of frames (or short clips). Given that *similarity* changes depending on its *viewpoint*, it is unrealistic to obtain frame-level similarity for all *viewpoints* in this manner.

This paper particularly focuses on video-level *similarities*. More concretely, we utilize the information of how multiple videos are divided into groups as an indicator of *similarity* because of its accessibility. For example, we have multiple video folders on our PCs or smart-phones, or we sometimes categorize videos on an Internet service. They are divided according to a reason, but in most cases, why they are grouped the way they are is unknown, or irrelevant to criteria, such as preference (liked or not liked). Thus, a *viewpoint* is not evident, but such video-level *similarity* can be measured as a mapping of *one viewpoint*.

In this paper, we assume the situation that multiple groups of videos that are divided based on *one similarity* are given, and we investigate how to introduce unknown underlying *viewpoint* to the summary. It is worth noting that, as we assume there are multiple possible ways to divide videos into groups depending on a *viewpoint* given the same set of videos, some overlap of content can exist between videos belonging to different groups, leading to technical difficulties, as we will state in Section 6.2.

For considering *similarity*, summaries extracted from similar videos should be similar, and ones extracted from different videos should be different from each other in addition to avoiding the redundancy derived from the original motivation of video summarization. In other words, given multiple groups of videos, the output summary of the video summarization

algorithm should be: (A) diverse, (B) representative of videos in the same group, and (C) discriminative against videos in the different groups.

To satisfy the requirements (A)-(C) simultaneously, we proposed a novel video summarization method from multiple groups of videos. Inspired by Fisher's discriminant criteria, it selects a summary by optimizing the combination of three terms the (a) inner-summary, (b) inner-group, and (c) between-group variance defined based on the feature representation of the summary, which can simply represent (A)-(C). In addition, we developed a novel optimization algorithm, which can be easily combined with feature learning, such as using convolutional neural networks (CNNs).

Moreover, we developed a novel dataset to investigate how well the generated summary reflects an underlying *viewpoint*. Because knowing individual *viewpoint* is generally impossible, we fixed it to two types of topics for each video. We also collected multiple videos that can be divided into groups based on these *viewpoints*. Quantitative and qualitative experiments were conducted on the dataset to demonstrate the effectiveness of proposed method.

The contributions of this paper are as follows:

- Propose a novel video summarization method from multiple groups of videos where their *similarity* are taken into consideration,
- Develop a novel dataset for quantitative evaluation
- Demonstrate the effectiveness of proposed method by quantitative and qualitative experiments on the dataset.

The remainder of this chapter is organized as follows. In Section 6.2, we discuss the related work of video summarization. Further, we explain the formulation and optimization of our video summarization method in Section 6.3. We state the detail of the dataset we created in Section 6.5, and describe and discuss the experiments that we performed on it in Section 6.6.

6.2 Methods for Video Summarization

Many recent studies have tackled the video summarization problem, and most of them can be categorized into either unsupervised or supervised approach. Unsupervised summarization [52, 46, 45, 47, 7, 18, 87, 30, 36, 37, 69, 50, 16] that creates a summary using specific selection criteria, has been conventionally studied. However, owing to the subjective property of this task, a supervised approach [42, 71, 59, 44, 25, 58, 26, 40, 21, 84], that trains a

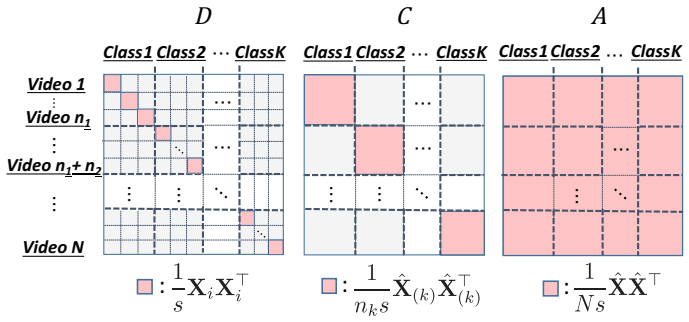


Figure 6.3: Overview of matrices D, C, and A, which are similarity matrices of inner-video, inner-group, and all videos. Non-zero elements of each matrix are colored pink and zero elements are colored gray.

summarization model which takes human-created summaries as the supervision, became standard because of its better performance. Most of their methods aim to extract one optimal summary and do not consider the *viewpoint*, which we focus on in this study.

The exception is query extractive summarization [64, 65] whose model takes a keyword as input and generates a summary based on it. It is similar to our work in that it assumes there can be multiple kinds of summaries for one video. However, our work is different in that we estimate what summary is created base on from the data instead of taking it as input. Besides, training model requires frame-level importance annotation for each keyword, which is unrealistic for real applications.

Some of the previous research worked on video summarization utilizing only other videos to alleviate the difficulty of building a dataset [10, 54, 55]. [10, 55] utilized other similar videos and aims to generate a summary that is (A) diverse, and (B) representative of videos in a similar group, but it is not considered to be (C) discriminative against videos in different groups. Given that not only *what is similar* but also *what is dissimilar* is essential to consider *similarity*, we attempt to generate a summary that meets all of the conditions, (A)-(C).

The research most relevant to ours is [54], which attempted to introduce discriminative information by utilizing a trained video classification model. It generates a summary with two steps. In the first step, it trains a spatio-temporal CNN that classifies the category of each video. In the second step, it calculates importance scores by spatially and temporally aggregating the gradients of the network's output with regard to the input over clips.

The success of this method has a strong dependence on the training in the first step. In this step, training is performed clip-by-clip by assigning the same label as that the video belongs to, to all clips of the video. Thus, it implicitly assumes all clips can be classified to the same group, and if there are some clips that are difficult to classify, it suffers from

over-fitting caused by trying to classify it correctly. Such a strong assumption does not apply in general, because generic videos (such as ones on YouTube) include various types of content. This assumption does not also apply in our case because we are interested in the situation where there are multiple possible ways to divide videos into groups given the same set of videos, as stated in Section 6.1, where some parts of videos can overlap with ones belonging to different groups for some *viewpoints*.

Unlike this, we do not assume all clips in the video can be classified correctly. Instead, our method considers the discrimination for only parts of videos. This makes it easy to find discriminative information even when there are visually similar clips across different groups.

We also acknowledge methods for discovering mid-level discriminative patches [68, 43, 31, 13–15] as related works because it attempts to find representative and discriminative elements from grouped data. Our work can be regarded as an extension of them to general videos.

6.3 Method

First, we introduce three quantities, that is, the (a) inner-summary, (b) within-group, and (c) between-group variances in subsection 6.3.1. Subsequently, we formulate our method by defining a loss function to meet the requirements discussed in Section 6.1. The optimization algorithm is described in subsection 6.3.2, and how to combine it with CNN feature learning is mentioned in subsection 6.3.3. The detailed derivation can be found in the supplemental material.

6.3.1 Formulation

Let $\mathbf{X}_i = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{T_i}]^\top \in \mathbb{R}^{T_i \times d}$ be a feature matrix for a video i with T_i segment (or frame) features $\mathbf{x}$. Our goal is to select s segments from the video. We start by defining the feature representation of the summary for video i as $\mathbf{v}_i = \frac{1}{s}\mathbf{X}_i^\top \mathbf{z}_i$, where $\mathbf{z}_i \in \{0,1\}^{T_i}$ is the indicator variable and $z_{it} = 1$ if the t-th segment is selected, and otherwise 0. It also has a constraint $||\mathbf{z}_i||_0 = s$ indicating that just s segments are selected as a summary. We can define a variance S_i^V for the summary of a video i as

$$S_i^V = \sum_{t=1}^{T_i} z_t (\mathbf{x}_t - \mathbf{v}_i)(\mathbf{x}_t - \mathbf{v}_i)^\top. \tag{6.1}$$

Algorithm 2 Optimization algorithm of (6.11)

0: **INPUT**: data matrix $Q = Q_1 - Q_2$, the number of selected clips s.
0: **INITIALIZE**: $\mathbf{z}_i = (1/s)\ \mathbf{1}_{T_i}$ for all video index i.
0: **repeat**
0: Calculate upper bound $\hat{L}_{(t)} = \hat{\mathbf{z}}^\top Q_1 \hat{\mathbf{z}} - 2\ \hat{\mathbf{z}}_{(t)}^\top Q_2 \hat{\mathbf{z}}$
0: Replace loss with $\hat{L}_{(t)}$ and solve QP problem.
0: **until** convergence
0: **RETURN** $\hat{\mathbf{z}}$ =0

Thus, its trace can be written as:

$$Tr(S_i^V) = \sum_{t=1}^{T_i} z_t \mathbf{x}_t^\top \mathbf{x}_t - \frac{1}{s} \mathbf{z}_i^\top \mathbf{X}_i \mathbf{X}_i^\top \mathbf{z}_i. \tag{6.2}$$

Placing all N videos together by using a stacked variable $\hat{\mathbf{z}} = [\mathbf{z}_1^\top, \mathbf{z}_2^\top, ..., \mathbf{z}_N^\top]^\top \in \{0,1\}^{\sum_{i=1}^N T_i}$, we can rewrite

$$Tr(S^V) = \sum_{i=1}^{N} Tr(S_i^V) = \hat{\mathbf{z}}^\top (F - D) \hat{\mathbf{z}}. \tag{6.3}$$

where F is a diagonal matrix whose element corresponds to $\mathbf{x}_t^\top \mathbf{x}_t$, and $D = \frac{1}{s} \oplus \sum_{i=1}^N \mathbf{X}_i \mathbf{X}_i^\top$ is a block diagonal matrix containing a similarity matrix of segments in the video i as i-th block elements.

By exploiting categorical information, we can also compute within-group variance S^W and between-group variance S^B. To compute them, we define the mean vector $\boldsymbol{\mu}_k$ for group $k \in \{1 : K\}$ and global mean vector $\bar{\boldsymbol{\mu}}$ as:

$$\boldsymbol{\mu}_k = \frac{1}{n_k} \sum_{i \in L_{(k)}} \mathbf{v}_i = \frac{1}{n_k s} \hat{\mathbf{X}}_{(k)}^\top \hat{\mathbf{z}}_{(k)}, \tag{6.4}$$

$$\bar{\boldsymbol{\mu}} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{v}_i = \frac{1}{Ns} \hat{\mathbf{X}}^\top \hat{\mathbf{z}}, \tag{6.5}$$

respectively. In these equations, $L_{(k)}$ is the set of indices of videos belonging to group k and $n_k = |L_{(k)}|$ (i.e., $N = \sum_k n_k$). In addition, $\hat{\mathbf{X}} = [\mathbf{X}_1^\top | \mathbf{X}_2^\top | ... | \mathbf{X}_N^\top]^\top \in \mathbb{R}^{(\sum_{i=1}^N T_i) \times d}$ is the matrix stacking all segment features of all videos. $\hat{\mathbf{X}}_{(k)}$ and $\hat{\mathbf{z}}_{(k)}$ are parts of $\hat{\mathbf{X}}$ and $\hat{\mathbf{z}}$, respectively, corresponding to videos contained by group k. We assume that a video index is ordered to satisfy $\hat{\mathbf{X}} = [\hat{\mathbf{X}}_{(1)}^\top | \hat{\mathbf{X}}_{(2)}^\top | ... | \hat{\mathbf{X}}_{(K)}^\top]^\top$. Here, the trace of within-group variance for

group k can be written as:

$$Tr(S^W_{(k)}) = Tr(\sum_{i \in L_{(k)}} s(\mathbf{v}_i - \boldsymbol{\mu}_k)(\mathbf{v}_i - \boldsymbol{\mu}_k)^\top)$$
$$= \frac{1}{s} \sum_{i \in L_{(k)}} \mathbf{z}_i^\top \mathbf{X}_i \mathbf{X}_i^\top \mathbf{z}_i - \frac{1}{n_k s} \hat{\mathbf{z}}_{(k)}^\top \hat{\mathbf{X}}_{(k)} \hat{\mathbf{X}}_{(k)}^\top \hat{\mathbf{z}}_{(k)}. \tag{6.6}$$

Aggregating them over all groups, the trace of within-group variance takes the following form:

$$Tr(S^W) = \sum_{k=1}^{K} Tr(S^W_{(k)}) = \hat{\mathbf{z}}^\top (D - C) \hat{\mathbf{z}}. \tag{6.7}$$

$C = \frac{1}{s} \oplus \sum_{k=1}^{K} \frac{1}{n_k} \hat{\mathbf{X}}_{(k)} \hat{\mathbf{X}}_{(k)}^\top$ is a block diagonal matrix containing a similarity matrix of segments in the video belonging to group k as a k-th block element. Similarly, the trace of between-group variance is:

$$\begin{aligned} Tr(S^B) &= Tr(\sum_{k=1}^{K} n_k s(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})^\top) \\ &= \hat{\mathbf{z}}^\top (C - A) \hat{\mathbf{z}}. \end{aligned} \tag{6.8}$$

In addition, matrix A is defined by $A = \frac{1}{Ns} \hat{\mathbf{X}} \hat{\mathbf{X}}^\top$. We show the overview of matrices D, C, and A in Fig. 6.3.

Loss function: We designed an optimization problem to meet the requirements discussed in Section 6.1: (A) diverse, (B) representative of videos in the same group, and (C) discriminative against videos in different groups. To simultaneously satisfy them, we minimized the within-group variance while maximizing the between-group and inner-video variances inspired by the concept of linear discriminant analysis. Thus, we maximized the following function, which is the weighted sum of the aforementioned three terms:

$$\lambda_1 Tr(S^V) - \lambda_2 Tr(S^W) + \lambda_3 Tr(S^B)$$
$$\text{s.t. } \lambda_1 \geq 0, \ \lambda_2 \geq 0, \ \lambda_3 \geq 0, \tag{6.9}$$

where λ_1, λ_2, λ_3 are hyper-parameters that control the importance of each term. We empirically fixed $\lambda_1 = 0.05$ in our experiments.

By substituting (6.3), (6.34), and (6.39) into (6.9), the optimization problem can be solved as:

$$\begin{aligned} \min \quad & \hat{\mathbf{z}}^{\top} Q \hat{\mathbf{z}} \\ Q \triangleq \quad & -\lambda_1 F + (\lambda_1 + \lambda_2) D - (\lambda_2 + \lambda_3) C + \lambda_3 A \\ \text{s.t.} \quad & ||\mathbf{z}_i||_0 = s,\ \mathbf{z}_i \in \{0,1\}^{T_i},\ \forall i \in \{1:N\} \end{aligned} \tag{6.10}$$

Table 6.1: The list of names for video groups (**target group**, **related group1**, **related group2**), and individual concepts of **target group** (**concept1**, **concept2**). We omit the article (e.g., the) before nouns due to the lack of space. We use the abbreviation of target group as [RV, RB, BS, DS, RD, SR, CC, RN, SC, RS] from top to bottom.

target group (TG)	concept1	concept2	related group1 (RG1)	related group2 (RG2)
running in Venice	Venice	running	running in Paris	shopping in Venice
riding bike on beach	beach	riding bike	riding bike in city	surfing on beach
boarding on snow mountain	snow mountain	boarding	boarding on dry sloop	hike in snow mountain
dog chasing sheep	sheep	dog	dog playing with kids	sheep grazing grass
racing in desert	desert	racing	racing in circuit	riding camel in desert
swimming and riding bike	swimming	riding bike	riding bike and tricking	diving and swimming
catching and cooking fish	catching fish	cooking fish	cooking fish in village	catching fish at river
riding helicopter in NewYork	NewYork	helicopter	riding helicopter in Hawaii	riding ship in NewYork
slackline and rock climbing	slackline	rock climbing	rock climbing and camping	slcakline and jaggling
riding horse in safari	safari	riding horse	riding horse in mountain	riding vehicle in safari

6.3.2 Optimization

Given that minimizing (6.10) directly is infeasible, we relaxed it to a continuous problem as follows:

$$\begin{aligned} &\min \quad \hat{\mathbf{z}}^\top Q \hat{\mathbf{z}} \\ &\text{s.t.} \quad P\hat{\mathbf{z}} = s\mathbf{1}_N,\ \hat{\mathbf{z}} \in [0,1]^{\sum_{i=1}^{N} T_i} \\ &\text{where} \quad P^\top = \begin{bmatrix} \mathbf{1}_{T_1} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{T_2} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{1}_{T_N} \end{bmatrix}. \end{aligned} \tag{6.11}$$

$\mathbf{1}_a$ indicates a vector whose elements are all ones and whose size is a, and the size of matrix P is $N \times \sum_{i=1}^{N} T_i$. The designed optimization problem is the difference of convex (DC) programming problem because all matrices that compose Q in (6.11) are positive semi-definite. We utilized a well-known CCCP (concave convex procedure) algorithm [81, 82] to solve it. Given the loss function represented by $L(\mathbf{x}) = f(\mathbf{x}) - g(\mathbf{x})$ where $f(\cdot)$ and $g(\cdot)$ are convex functions, the algorithm iteratively minimizes the upper bound of loss calculated by the linear approximation of $g(\mathbf{x})$. Formally, in the iteration t, it minimizes: $\hat{L}(\mathbf{x}) = f(\mathbf{x}) - \partial_{\mathbf{x}} g(\mathbf{x}_{(t)})^\top \mathbf{x} \geq L(\mathbf{x})$. In our problem, the loss function can be decomposed into the difference of two convex functions: $\hat{\mathbf{z}}^\top Q \hat{\mathbf{z}} = \hat{\mathbf{z}}^\top Q_1 \hat{\mathbf{z}} - \hat{\mathbf{z}}^\top Q_2 \hat{\mathbf{z}}$, where $Q_1 \triangleq (\lambda_1 + \lambda_2)D + \lambda_3 A$ and $Q_2 \triangleq \lambda_1 F + (\lambda_2 + \lambda_3)C$. We optimized the following quadratic programming (QP) problem in t-th iteration,

$$\begin{aligned} &\min \quad \hat{\mathbf{z}}^\top Q_1 \hat{\mathbf{z}} - 2\, \hat{\mathbf{z}}_{(t)}^\top Q_2 \hat{\mathbf{z}} \\ &\text{s.t.} \quad P\hat{\mathbf{z}} = s\mathbf{1}_N,\ \hat{\mathbf{z}} \in [0,1]^{\sum_{i=1}^{N} T_i}, \end{aligned} \tag{6.12}$$

where $\hat{\mathbf{z}}_{(t)}$ is the estimation of $\hat{\mathbf{z}}$ in the t-th iteration. In our implementation, we used a CVX package [23, 22] to solve the QP problem (6.12). An overview of our algorithm is shown in Algorithm 2. Please refer [41] for the convergence property of CCCP.

6.3.3 Feature learning

To obtain the feature representation that is more suitable for video summarization, feature learning is applied. Firstly, we replace the visual feature $\mathbf{x}$ in subsection 6.3.1 to $f(\mathbf{x}; \mathbf{w})$ where $f(\cdot)$ is a feature extractor function that is differentiable with regard to the parameter $\mathbf{w}$ and the input $\mathbf{x}$ is a sequence of raw frames in the RGB space. Specifically, we exploited the

C3D network [74] as a feature extractor. Fixing $\hat{\mathbf{z}}$, the loss function (6.11) can be written as:

$$L = \sum_{i,j} \hat{z}_i \hat{z}_j m_{ij} f(\mathbf{x}_i)^\top f(\mathbf{x}_j), \tag{6.13}$$

where $\hat{z}_i$ is i-th element of $\hat{\mathbf{z}}$. Also, m_{ij} is the ij-th element of matrix M written as follows:

$$M = -\lambda_1 \mathbf{1}_F + (\lambda_1 + \lambda_2)\mathbf{1}_D - (\lambda_2 + \lambda_3)\mathbf{1}_C + \lambda_3 \mathbf{1}_A.$$

Here, $\mathbf{1}_X$ represents an indicator matrix whose element takes 1 where the corresponding element of X is not 0, and takes 0 otherwise. We optimize the loss function with regard to the parameter by stochastic gradient decent (SGD). Because many of $\hat{z}_i$ are small values or zeros, minimizing (6.13) directly is not efficient. We avoid the inefficiency by sampling samples $\mathbf{x}_i$ based on their weight $\hat{z}_i$. Given $\sum \hat{z}_i = Ns$, we sample $\mathbf{x}_i$ from the distribution $p(\mathbf{x}_i) = \hat{z}_i / Ns$ (≥ 0) and stochastically minimize the expectation:

$$\mathbb{E}_{\mathbf{x}_i, \mathbf{x}_j \sim p(\mathbf{x})}[m_{ij} f(\mathbf{x}_i)^\top f(\mathbf{x}_j)]. \tag{6.14}$$

In an iteration when updating parameters, the model fetches pairs $(\mathbf{x}_i, \mathbf{x}_j)$ and computes the dot product of the feature representations. The loss for this batch is calculated by summing up the dot product weighted by m_{ij}. We repeatedly and alternately compute the summary via the Algorithm 2 and optimize the parameter of the feature extractor.

6.4 Relationship with Mutual Information

Proposed method can be seen as a special case of mutual information maximization. We assume underlying distribution as an isotropic gaussian-distribution. Formally, with positive constant values α and β,

$$\mathbf{c} \sim \mathcal{N}(\boldsymbol{\mu}, \frac{\alpha^{-1}}{2} I) \tag{6.15}$$

$$\mathbf{v} \mid \mathbf{c} \sim \mathcal{N}(\mathbf{c}, \frac{\beta^{-1}}{2} I) \tag{6.16}$$

$$\mathbf{x} \mid \mathbf{v} \sim \mathcal{N}(\mathbf{v}, \frac{\gamma^{-1}}{2} I) \tag{6.17}$$

Table 6.2: statistics of dataset

group	# of videos	# of frames	duration
TG	50	243,873	8,832(s)
RG1 + RG2	100	440,330	15,683(s)

From the relationship of marginal distribution of gaussian distribution, we obtain

$$p(\mathbf{v}) = \mathcal{N}(\boldsymbol{\mu}, \frac{\rho^{-1}}{2} I) \tag{6.18}$$

$$p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \frac{\sigma^{-1}}{2} I) \tag{6.19}$$

where $\rho^{-1} = \alpha^{-1} + \beta^{-1}$, $\sigma^{-1} = \alpha^{-1} + \beta^{-1} + \gamma^{-1}$, and $\tau^{-1} = \beta^{-1} + \gamma^{-1}$
Given observed variables $\mathbf{x}$, an unbiased estimator of the expectation of above three distributions are obtained by simply taking mean because they are gaussian. In our case, we have s shot features for each video, n_k videos for each group, and K groups in total. Thus, they can be estimated by $\hat{\mathbf{v}}_j = \frac{1}{s}\sum_{i=1}^{s} \mathbf{x}_i$, $\hat{\mathbf{c}}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} \hat{\mathbf{v}}_j$, $\hat{\boldsymbol{\mu}} = \frac{1}{K}\sum_{k=1}^{K} \hat{\mathbf{c}}_k$.

The mutual information of two random variables $\mathbf{c}, \mathbf{v}$ are

$$\begin{aligned}
\mathrm{MI}(\mathbf{c},\mathbf{x}) &= \mathbb{E}_{\mathbf{c},\mathbf{x}}\left[\log \frac{p(\mathbf{x}|\mathbf{c})}{p(\mathbf{x})}\right] && (6.20)\\
&= \mathbb{E}_{\mathbf{c},\mathbf{x}}\left[\log \frac{\exp(-\tau||\mathbf{x}-\mathbf{c}||^2)}{\exp(-\sigma||\mathbf{x}-\boldsymbol{\mu}||^2)}\right] + \text{const} && (6.21)\\
&= \mathbb{E}_{\mathbf{c},\mathbf{x}}\left[\sigma||\mathbf{x}-\boldsymbol{\mu}||^2 - \tau||\mathbf{x}-\mathbf{c}||^2\right] && (6.22)\\
& && (6.23)
\end{aligned}$$

Terms that data selection affects is empirically estimated.

$$\sigma(Tr(S^{(V)}) + Tr(S^{(W)}) + Tr(S^{(B)})) - \tau(Tr(S^{(V)}) + Tr(S^{(W)}))) \tag{6.24}$$

$$(\sigma-\tau)Tr(S^{(V)}) + (\sigma-\tau)Tr(S^{(W)}) + \sigma Tr(S^{(B)}) \tag{6.25}$$

$$-\frac{\tau^2}{\alpha\tau}Tr(S^{(V)}) - \frac{\tau^2}{\alpha\tau}Tr(S^{(W)}) + \sigma Tr(S^{(B)}) \tag{6.26}$$

With appropriate resetting hyper-parameters, the problem of jointly maximizing two mutual informations comes down to

$$\mathrm{MI}(\mathbf{c},\mathbf{x}) = \lambda_1 Tr(S^{(V)}) - \lambda_2 Tr(S^{(W)}) + \lambda_3 Tr(S^{(B)}), \tag{6.27}$$

which is coincident with optimization problem in the previous section.

(a) safari (above) and riding horse (below)

(b) slackline (above) and rock climbing (below)

(c) NewYork (above) and riding helicopter (below)

(d) catching fish (above) and cooking fish (below)

Figure 6.4: Example human-created summary of video whose **target group** are "riding horse in safari" (upper left), "slackline and rock climbing" (upper right), "riding helicopter in NewYork" (lower left), and "catching and cooking fish" (lower right) based on the concept written in each figure.

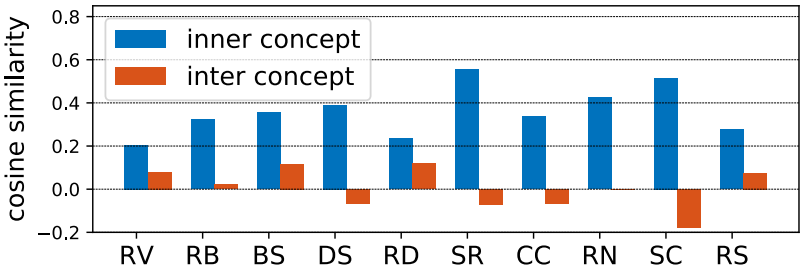


Figure 6.5: Mean cosine similarity of human-assigned scores for each target group. We denote the value computed from the score pairs that are assigned to the same concept and different concepts as inner concepts (blue) and inter concepts (orange), respectively. When referring to the abbreviated names of groups, please refer to the Table 6.1.

6.5 Dataset

The motivation of this study is the claim that an optimal summary should be varied depending on a *viewpoint*, and this paper deals with this by considering the *similarities*. To investigate how well the underlying *viewpoint* are taken into consideration, given multiple groups of videos that are divided based on the *similarity*, we compiled a novel video summarization dataset[2]. Quantitative evaluation is challenging because the *viewpoint* is generally unknown. Thus, for the purpose of quantitative evaluation, we collected a set of videos that can have two interpretable ways of separation assuming they have corresponding *viewpoint*. In addition,

[2]Dataset is available at https://akanehira.github.io/viewpoint/.

we collected human-created summaries fixing the importance criteria to two concepts based on each *viewpoint*. The procedure of building the dataset is as follows:

First, we collected five videos that match the topics written in **target group (TG)**, **related group1 (RG1)**, **related group2 (RG2)** of Table 6.1 by retrieving them in YouTube[3] using a keyword. Each of **TG**, **RG1**, **RG2** has two explicit concepts such that they can be visually confirmed; e.g., location, activity, object, and scene. The concepts of **TG** are written in **concept1** and **concept2** columns in the table, and both **RG1** and **RG2** were chosen to share either one of them. There are two interpretable ways to divide these sets of videos, i.e., (**TG + RG1**) vs. (**RG2**) and (**TG + RG2**) vs. (**RG1**) because **RG1** and **RG2** share one topic with **TG**. Assuming these divisions are based on one *viewpoint*, we collected the summary based on it using two concepts for videos belonging to **TG**. For example, if we are given two groups, one of which contains "running in Venice" and "running in Paris" videos, and the other group includes "shopping in Venice" videos, the underlying *viewpoint* is expected to be "kind of activity." For such a scenario, we collected summaries based on "running" for the videos of "running in Venice."

For annotating the importance of each frame of the video belonging to **TG**, we used Amazon Mechanical Turk (AMT). Firstly, videos were evenly divided into clips beforehand so that the length of each clip was two seconds long following the setting of [69]. Subsequently, after workers watched a whole video, they were asked to assign a importance score to each clip of the video, assuming that they created a summary based on a pre-determined topic, which corresponds to the concept written in **concept1** or **concept2** columns in the Table 6.1. Importance scores are chosen from 1 (not important) to 3 (very important), and workers were asked to guarantee the number of clips having a score of 3 falls in the range between 10% and 20% of the total number of clips in the video. For each video and each **concept**, five workers were assigned.

We display the statistics of the dataset and some example of the human-created summary in Table 6.2 and Fig. 6.4, respectively. Also, in order to investigate how similar the assigned score between subjects is, we calculated the similarity of the score vector. After subtracting the mean value from each score, the mean cosine similarity for the pair of scores that are assigned for the same concepts (e.g., **concept1** and **concept1**) and different concepts (**concept1** and **concept2**) were separately computed, and the result is shown in Fig. 6.5. As we can see in the table, the similarity of scores that comes from the inner-concept is higher than that of inter-concept, which indicates that the importance depends on the *viewpoint* of the videos.

[3]https://www.youtube.com/

6.6 Experiment

6.6.1 Preprocessing

To compute the segment used as the smallest element for video summarization, we followed a simple method proposed in [10]. After counting the difference of two consecutive frames in the RGB and HSV space, the points on which the total amount of change exceeds 75% of all pixels were regarded as change points. Subsequently, we combined short clips into the following clip and evenly divided the long clips in order such that the number of frames in each clip was more than 32 and less than 112.

6.6.2 Visual features

For obtaining frame-level visual features, we exploited the intermediate state of the C3D [74] network, which is known to be so generic that it can be used for other tasks, including video summarization [55]. We extracted the features from an fc6 layer of a network pre-trained on a Sports1M [35] dataset. The length of the input was 16 frames, and features were extracted every 16 frames. The dimension of the output feature vector was 4,096. Clip-level representations were calculated by performing an average pooling over all frame-level features in each clip followed by a l_2 normalization.

6.6.3 Evaluation

For a quantitative evaluation, we compared automatically generated summaries with human made ones. First, we explain the grouping setting of videos. There are two interpretable ways of grouping that include each target group as stated in Section 6.5:

- regarding **related group2 (RG2)** as the same group as **target group (TG)** and **related group1 (RG1)** as the different group (setting1).
- regarding **related group1 (RG1)** as the same group as **target group (TG)** and **related group2 (RG2)** as the different group (setting2).

In the case that the grouping setting1 was used, we evaluated it with the summary annotated for **concept1**. Alternatively, when videos are divided like setting2, the summary for **concept2** was used for the evaluation. Note we treated each **TG** independently in throughout this experiment.

We set the ground-truth summary in the following procedure. The mean of the importance scores were calculated over all frames in each clip, which was determined by the method described in the previous subsection. The top-30% of the number of all clips whose importance

scores are highest were extracted from each video and regarded as ground-truth. As an evaluation metric, we computed the mean Average Precision (MAP) from a pair of summaries, and reported the mean value. Formally, for each **TG**, $1/(CIJ)\sum_{c=1}^{C}\sum_{j=1}^{J}\sum_{i=1}^{I} AP(l_{(c)}^{ij}, \hat{l}_{(c)}^{i})$ was calculated where l and $\hat{l}$ are ground-truth summaries and the predicted summary, respectively. C indicates the number of concepts on which the summary created by the annotators is based on. I, J are the number of subjects and the number of videos in the group respectively. In particular, (C, I, J) were (2, 5, 5) as written in Section 6.5 in this study.

6.6.4 Implementation detail

As stated in Section 6.3, we used a C3D network [74] pre-trained on a Sports1M dataset [35], which has eight convolution layers followed by three fully connected layers. During fine-tuning, the initial learning rate was 10^{-5}. Weight decay and momentum were set to 10^{-4} and 0.9 respectively. The number of repetitions of the feature learning and summary estimation was set to 5. The number of epochs for each repetition was 10, and the learning rate was multiplied by 0.9 for every epoch. Here, epoch indicates {# of all clips}/{batch size} iteration even though clips were not uniformly sampled.

6.6.5 Comparison with other methods

To investigate the effectiveness of the proposed method, we compared it with other baseline methods as follows:

Sparse Modeling Representative Selection (SMRS) [17]: SMRS computes a representation of video clips such that a small number of clips can represent an entire video by group sparse regularization. We selected clips whose l_2 norm of representation was the largest.

kmeans (CK) and spectral clustering (CS): One simple solution to extract representative information between multiple videos is applying clustering algorithm. We applied two clustering algorithms, namely kmeans (CK) and spectral clustering (CS), for all clips of video which was regarded as the same groups. RBF kernel was used to build an affinity matrix necessary for computation of spectral clustering. The number of clusters was set to 20 as in [54]. Summaries were generated by selecting clips that are the closest to the cluster center of the largest clusters.

Maximum Bi-Clique Finding (MBF) [10]: The MBF is a video co-summarization algorithm that extracts a bi-clique from a bi-partite graph with a maximum inner weight. MBF algorithms were applied to each pair of videos within a video group, and the quality

scores were computed by aggregating the results of all pairs. We used hyper-parameters same as the ones suggested in the original paper [10].

Collaborative Video Summarization (CVS) [55]: CVS is the method that computes a representation of a video clip based on sparse modeling, similar to SMRS. The main difference is that CVS aims to extract a summary that is representative of other videos belonging to the same group as well as the video. We selected the clips whose l_2 norm of representation was the largest. The decision of hyper-parameters follows the original paper [55].

Weakly Supervised Video Summarization (WSVS) [54] : Similar to our method, WSVS creates a summary using multiple groups. It computes the importance score by calculating the gradient of the classification network with regard to the input space, and aggregating it over a clip. The techniques for training the classification network such as network structure, learning setting, and data augmentation, followed the original paper [54]. For a fair comparison, we leveraged the same network as the one we used as well as the one proposed in the original paper pre-trained on split-1 of the UCF101 [70] dataset (denoted as WSVS (large) and WSVS respectively). Moreover, all clips were used for training, and gradients were calculated for them.

The top-5 and top-10 MAP are shown in Table 6.3. First, our method performed better than the other methods, which consider only the representativeness from a single group, in most of the **target groups**, and showed competitive performance in the other. It implies that discriminative information is the key to estimating the *viewpoint*.

Secondly, the performance of our methods with feature learning was better than that without it as a whole. We found it works well even though we exploited a large network with enormous parameters and the number of samples was relatively small in many cases, except in a few categories. When considering "riding bike on beach (RB)" or "boarding on a snow mountain (BS)", we noticed a drop in the performance. Our feature learning algorithm works in a kind of self-supervised manner; It trains the feature extractor to explain the current summary better, and therefore, it is dependent on the initial summary selection. If outliers have a high importance score in that step, no matter whether it is discriminative, the parameter update is likely to be strongly affected by such outliers, which causes a performance drop.

Thirdly, we found the performance of WSVS and WSVS (large) were worse than our method and even than CSV, which uses only one group. We assume the reason is that it failed to train the classification model. This method trains the classification model clip-by-clip by assigning the same label to all video clips. It implicitly assumes all clips can be classified into the same group, which is unrealistic when using generic videos such as ones on the web as stated in Section 6.2. If there are some clips that are difficult or impossible to classify, it

Table 6.3: Top-5 mean AP computed from human-created summary and predicted summary for each method. Results are shown for each **target group**. For referring to the abbreviated names of groups, please see the Table 6.1.

	RV	RB	BS	DS	RD	SR	CC	RN	SC	RS	mean
SMRS [17]	0.318	0.371	0.338	0.314	0.283	0.317	0.294	0.348	0.348	0.286	0.322
CK	0.329	0.321	0.291	0.269	0.318	0.271	0.275	0.295	0.305	0.268	0.294
CS	0.318	0.330	0.309	0.317	0.278	0.293	0.302	**0.355**	0.350	0.271	0.312
MBF [10]	**0.387**	0.332	0.345	0.316	0.319	0.324	0.375	0.317	0.324	0.288	0.333
CVS [55]	0.339	0.365	**0.388**	0.334	0.359	0.386	0.362	0.303	0.337	0.356	0.353
WSVS [54]	0.333	0.339	0.310	0.331	0.272	0.335	0.336	0.303	0.329	0.330	0.322
WSVS (large) [54]	0.331	0.350	0.322	0.294	0.304	0.306	0.308	0.322	0.342	0.310	0.319
ours	0.373	**0.382**	0.367	0.396	0.327	0.497	0.374	0.340	0.368	0.368	0.379
ours (feature learning)	0.372	0.376	0.299	**0.403**	**0.373**	**0.518**	**0.388**	0.338	**0.408**	**0.378**	**0.385**

suffers from over-fitting caused by attempting to correctly classify them. In our case, we assume there are multiple possible ways to divide videos into groups given the same set of

Table 6.4: Top-10 mean AP computed from human-created summary and predicted summary for each method. Results are shown for each **target group**. For referring to the abbreviated names of groups, please see the Table 6.1.

	RV	RB	BS	DS	RD	SR	CC	RN	SC	RS	mean
SMRS [17]	0.354	0.370	0.373	0.335	0.320	0.344	0.309	0.374	0.365	0.344	0.349
CK	0.386	0.334	0.393	0.295	0.337	0.280	0.335	**0.434**	0.400	0.289	0.349
CS	0.344	0.344	0.326	0.333	0.319	0.304	0.330	0.384	**0.450**	0.310	0.344
MBF [10]	0.402	0.362	0.352	0.372	0.355	0.314	0.352	0.416	0.354	0.331	0.361
CVS [55]	0.370	0.382	**0.404**	0.381	0.358	0.387	0.374	0.376	0.408	0.380	0.382
WSVS [54]	0.358	0.303	0.356	0.353	0.318	0.368	0.359	0.344	0.349	0.323	0.343
WSVS (large) [54]	0.372	0.333	0.365	0.350	0.322	0.319	0.343	0.343	0.384	0.319	0.345
ours	**0.404**	0.393	0.366	0.423	0.338	0.540	0.412	0.386	0.387	0.375	0.402
ours (feature learning)	0.395	**0.407**	0.335	**0.430**	**0.363**	**0.545**	**0.423**	0.375	0.399	**0.393**	**0.406**

videos, as stated earlier. Therefore, parameters cannot be appropriately learned because some clips in videos belonging to different groups can appear to be similar. Given that our method

Table 6.5: User study results for the quality evaluation.

method	MBF [10]	CVS [55]	ours
score	1.07	1.22	**1.32**

Table 6.6: User study results for topic selection task. The accuracy takes the value in the range $[0, 1]$.

method	MBF [10]	CVS [55]	ours
accuracy	0.47	0.60	**0.76**

considers the discrimination of the generated summary, not all clips, it worked better even when using CNN with large parameters.

6.6.6 User study

Because video summarization is a relatively subjective task, we also evaluated the performance with a user study. We asked crowd-workers to assign the quality score to summaries generated from MBF, CVS, and proposed method. They chose the score from -2 (bad) to 2 (good), and for each video and concept, 10 workers were assigned. The mean results are shown in Table 6.5. It indicates that the quality of summaries of our method is the best among three methods.

6.6.7 Visualizing the reason of group division

One possible application of our method is visualizing the reason driving group divisions. Given multiple groups of videos, why they are grouped in such way is unknown, our algorithm works to visualize an underlying visual concept that is a criterion of the division. To determine how well our algorithm has the ability of this, we performed a qualitative evaluation using AMT. We asked crowd-workers to select the topic out of either **concept1** or **concept2** for summaries created in the group setting1 and setting2. We evaluated the performance of how well workers can answer questions about a topic correctly. We set the ground-truth topic as **concept1** when setting1 was used and **concept2** for setting2. We assigned 10 workers for each summary and each setting. As shown in the Table 6.6, our method performed better than other methods, which indicates the ability to explain the reason behind grouping.

6.7 Relationship with other methods

The maximum bi-clique finding (MBF) technique [10] for video co-summarization builds a bi-partite graph for two videos, on which each segment corresponds to a node. Let $\mathbf{u} \in \{0,1\}^N, \mathbf{v} \in \{0,1\}^M$ be a vector indicating a selection of segments from video U and V, and $C \in \mathbb{R}^{N \times M}$ be the similarity matrix between the segments of two videos used as an edge weight. This method finds a bi-clique from the graph with the maximum summation of weight. Formally, it maximizes $\mathbf{u}^T C \mathbf{v}$ by using the constraint $u_i + v_j \leq 1 + I(C_{ij} \geq \varepsilon)$, where the indicator is $I(\cdot) = 1$ when the condition is met, otherwise it is 0, and ε is the predefined threshold value.

The connection between this and our proposed methods can be observed. If we set $\lambda_3 = 0$ in (10) in the main paper by ignoring the videos in other groups and assume that we treat only two samples (i.e., $n_k = 2$) denoting their selection vector as $\mathbf{u} \in \{0,1\}^N, \mathbf{v} \in \{0,1\}^M$, the optimization problem in (10) in the main paper can be rewritten as

$$\max \begin{bmatrix} \mathbf{u}^T & \mathbf{v}^T \end{bmatrix} \begin{bmatrix} -(\lambda_1 - \frac{1}{2}\lambda_2)K_{UU} & \frac{1}{4}\lambda_2 K_{UV} \\ \frac{1}{4}\lambda_2 K_{UV}^\top & -(\lambda_1 - \frac{1}{2}\lambda_2)K_{VV} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix}$$

where K_{UU}, K_{UV}, K_{VV} indicate the kernel matrices of shots features in the video U and U, U and V, and V and V respectively. (In this paper, we utilized linear kernel instead of rbf kernel used in [10].) For simplicity, we assume features are normalized to meet $k(\mathbf{x}, \mathbf{x}) = \mathbf{1}$ for all shot features $\mathbf{x}$. If we set $\lambda_2 = 2\lambda_1$, the block diagonal matrix will become 0, and the problem is simplified to the selection of a set of nodes from a bi-partite graph with the maximum inner weight, corresponding to $\varepsilon = 0$ in the MBF technique. From this, our algorithm can be regarded as a kind of generalization of MBF algorithm.

Furthermore, by only considering the first term (i.e., $(\lambda_2 = 0, \lambda_3 = 0)$), we can find an analogy to methods that aim to preserve diversity. For example, the DPP [40] extracts a subset whose determinant of the kernel matrix is the maximum, and Lu et al. [46] aims to minimize the similarity of consecutive frames in the summary. Our approach is different in that it minimizes the summation of all similarities in the summary, but it shares the same motivation as them.

6.8 Further results of user study

In the main paper, we fixed the *viewpoint*, and we compared the generated summaries with the ones created based on one explicit concept, which can be expressed with a few words, due to the difficulty of quantitative evaluation. We also conducted user study that measures

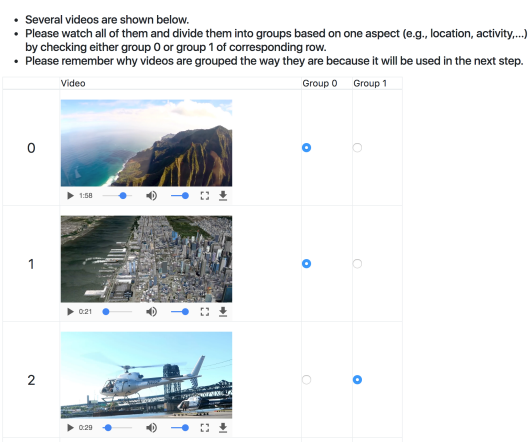
the ability to estimate underlying *viewpoint* with weaker constraint using the same dataset. For this purpose, we developed AMT-like web page as shown in Fig. 6.6a and Fig. 6.6b.

Firstly, four videos were randomly picked from each of **TG**, **RG1**, **RG2**, and they were shown to the subjects. Subjects were asked to split them into two groups based on one criterion which they decided on their own. Subsequently, they watched summaries of those videos belonging to **TG** generated by MBF [10], CVS [55], and ours (without feature learning). The summary which most reflects the criterion that was used to divide videos into groups was selected. (It was allowed to choose multiple summaries. Moreover, if there were not appropriate one, subjects do not need to choose anything.) For each task, five workers were assigned.

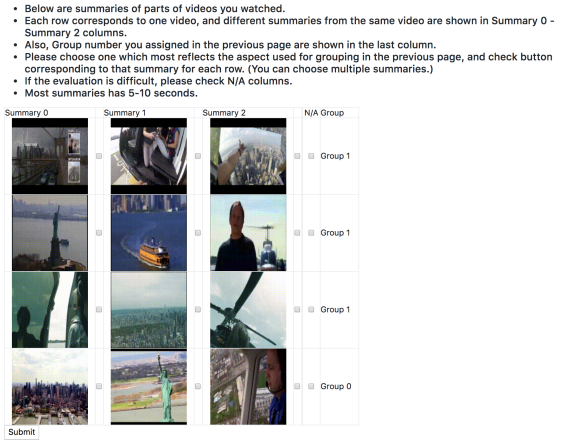
Table 6.7: The ratio that the summary generated from each method were selected. N/A means no method were selected.

	N/A	MBF [10]	CVS [55]	ours
score	0.09	0.37	0.38	**0.50**

We show the number that each method were selected divided by the number of videos in the Table 6.7, and the score of our method is better than the others in it. This result indicates that our method can generate the summary that explains the criteria of grouping when the *viewpoint* changes person to person.



(a) The screenshot image of the web page used for dividing videos to groups.



(b) The screenshot image of the web page used for the evaluation of summaries.

Figure 6.6: The screenshot of web pages developed for the user study evaluation.

6.9 Detailed result of topic selection task

Per-group accuracy of the topic selection task in the subsection 5.7 are displayed in the Fig. 6.7. We can see the topic of the summary generated by our algorithm is correctly answered with higher probability than other methods, which demonstrates the ability to recover the criteria of grouping. The performance of MBF was near random rate (0.5), and worse than that in several groups. We conjecture the reason attributes to the fact that MBF uses only two videos to find the visual co-occurrence. If the feature representation of shots which is representative to topics are similar each other, it may fail to find the common pattern within the group.

6.10 Additional Analysis

Applicability for long videos: To investigate the applicability of the proposed method to long videos, 2 out of 5 videos in each group were expanded to 5 times longer by synthesizing it with randomly selected clips in other irrelevant videos and set their scores to 0. The top-5 mAP of MBF, CSV, and ours got 0.217, 0.221, and 0.275 respectively. Results showed the applicability of proposed algorithm for long videos.

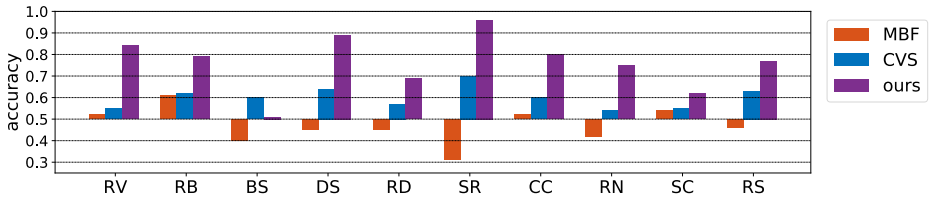


Figure 6.7: Per-group accuracy of topic selection task. Each bar corresponds to the each method, namely, MBF [10] (orange), CVS [55] (blue), and ours (purple). Please note 0.5 (random rate) are set to the center of this graph. For referring to the abbreviated names of groups, please see the Table 1 in the main paper.

Comparison with human performance: We also compared the performance with that of the summary created by human. Treating a summary for one user as a prediction, we computed mAP in the same way with the main experiment, and we regarded the human performance by averaging them. The average score of the human summary was 0.456, and 0.498 respectively. The performance of our method was approximately 80% compared with it.

Computation time: Average computation time per video of MBF, CVS, ours, and ours (feature learning) are 0.02(s), 36.82(s), 42.82(s), and 3562.34(s) with 1 CPU (Intel Xeon, 2.60GHz) and 2 GPUs (Tesla K40).

Ablation study: The top-5 mAP when dropping λ_1, λ_2, λ_3, and nothing are 0.370, 0.365, 0.336, and 0.379, which reveals the importance of discriminativeness.

Choosing hyper-parameters and their sensitivity: Fixing λ_3 to 1.0, and empirically setting λ_1 to 0.05, we changed λ_2 in [0.0, 1.0] at 0.1 interval. and we found performance is not sensitive to λ_2 unless it reaches to 0.0 or 1.0. For fair comparison, we showed the performance of best parameter ($\lambda_2 = 0.1$) in the same way as other methods.

6.11 Detailed derivation of equations

6.11.1 Trace of inner-video variance

$$
\begin{aligned}
\mathrm{Tr}(S_i^V) &= \mathrm{Tr}(\sum_{t=1}^{T_i} z_t(\mathbf{x}_t - \mathbf{v}_i)(\mathbf{x}_t - \mathbf{v}_i)^\top) && (6.28)\\
&= \sum_{t=1}^{T_i} \mathrm{Tr}(z_t(\mathbf{x}_t - \mathbf{v}_i)(\mathbf{x}_t - \mathbf{v}_i)^\top) && (6.29)\\
&= \sum_{t=1}^{T_i} z_t(\mathbf{x}_t - \mathbf{v}_i)^\top(\mathbf{x}_t - \mathbf{v}_i) && (6.30)\\
&= \sum_{t=1}^{T_i} z_t\mathbf{x}_t^\top\mathbf{x}_t - 2\mathbf{v}_i^\top\sum_{t=1}^{T_i} z_t\mathbf{x}_t + \sum_{t=1}^{T_i} z_t\mathbf{v}_i^\top\mathbf{v}_i && (6.31)\\
&= \sum_{t=1}^{T_i} z_t\mathbf{x}_t^\top\mathbf{x}_t - \frac{2}{s}\mathbf{z}_i^\top\mathbf{X}_i\mathbf{X}_i^\top\mathbf{z}_i + \frac{1}{s}\mathbf{z}_i^\top\mathbf{X}_i\mathbf{X}_i^\top\mathbf{z}_i && (6.32)\\
&= \sum_{t=1}^{T_i} z_t\mathbf{x}_t^\top\mathbf{x}_t - \frac{1}{s}\mathbf{z}_i^\top\mathbf{X}_i\mathbf{X}_i^\top\mathbf{z}_i && (6.33)
\end{aligned}
$$

(6.29) and (6.30) are derived by an identity $\mathrm{Tr}(\sum_i A_i) = \sum_i \mathrm{Tr}(A_i)$, and $\mathrm{Tr}(\mathbf{a}\mathbf{a}^\top) = \mathbf{a}^\top\mathbf{a}$. To derive (6.32), we utilize the definition $\mathbf{v}_i = \frac{1}{s}\mathbf{X}_i^\top\mathbf{z}_i$ and constraint $||\mathbf{z}_i||_0 = s$.

6.11.2 Trace of within-class variance

$$
\begin{aligned}
\mathrm{Tr}(S_{(k)}^W) &= \mathrm{Tr}(\sum_{i\in L_{(k)}} s(\mathbf{v}_i - \boldsymbol{\mu}_k)(\mathbf{v}_i - \boldsymbol{\mu}_k)^\top) && (6.34)\\
&= \sum_{i\in L_{(k)}} s(\mathbf{v}_i - \boldsymbol{\mu}_k)^\top(\mathbf{v}_i - \boldsymbol{\mu}_k) && (6.35)\\
&= s\sum_{i\in L_{(k)}} \mathbf{v}_i^\top\mathbf{v}_i - 2s(\sum\mathbf{v}_i)^\top\boldsymbol{\mu}_k + n_k s\boldsymbol{\mu}_k^\top\boldsymbol{\mu}_k && (6.36)\\
&= \frac{1}{s}\sum_{i\in L_{(k)}} \mathbf{z}_i^\top\mathbf{X}_i\mathbf{X}_i^\top\mathbf{z}_i - \frac{2}{n_k s}\hat{\mathbf{z}}_{(k)}^\top\hat{\mathbf{X}}_{(k)}\hat{\mathbf{X}}_{(k)}^\top\hat{\mathbf{z}}_{(k)} + \frac{1}{n_k s}\hat{\mathbf{z}}_{(k)}^\top\hat{\mathbf{X}}_{(k)}\hat{\mathbf{X}}_{(k)}^\top\hat{\mathbf{z}}_{(k)} && (6.37)\\
&= \frac{1}{s}\sum_{i\in L_{(k)}} \mathbf{z}_i^\top\mathbf{X}_i\mathbf{X}_i^\top\mathbf{z}_i - \frac{1}{n_k s}\hat{\mathbf{z}}_{(k)}^\top\hat{\mathbf{X}}_{(k)}\hat{\mathbf{X}}_{(k)}^\top\hat{\mathbf{z}}_{(k)} && (6.38)
\end{aligned}
$$

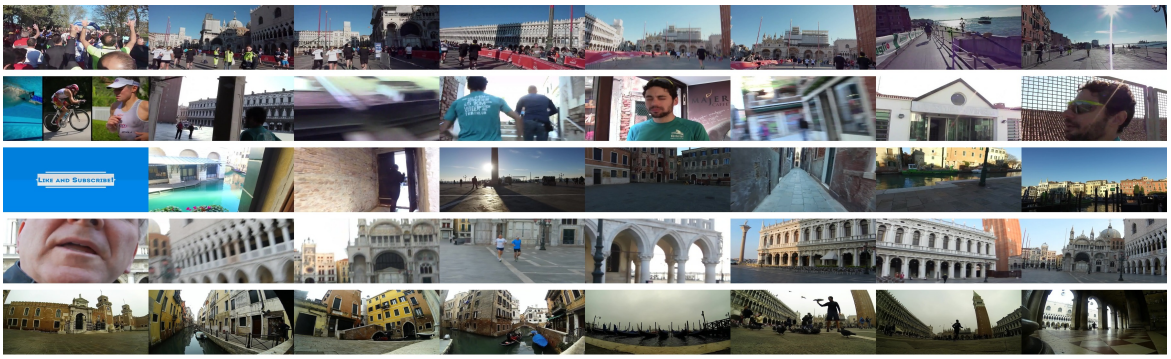
6.11.3 Trace of between-class variance

$$
\begin{aligned}
\mathrm{Tr}(S^B) &= \mathrm{Tr}(\sum_{k=1}^{K} n_k s(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})^\top) \\
&= \sum_{k=1}^{K} n_k s(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}})^\top(\boldsymbol{\mu}_k - \bar{\boldsymbol{\mu}}) && (6.39) \\
&= s\sum_{k=1}^{K} n_k \boldsymbol{\mu}_k^\top \boldsymbol{\mu}_k - 2s\bar{\boldsymbol{\mu}}^\top(\sum_{k=1}^{K} n_k \boldsymbol{\mu}_k) + Ns\bar{\boldsymbol{\mu}}^\top\bar{\boldsymbol{\mu}} && (6.40) \\
&= \frac{1}{s}\sum_{k=1}^{K}\frac{1}{n_k}\hat{\mathbf{z}}_{(k)}^\top\hat{\mathbf{X}}_{(k)}\hat{\mathbf{X}}_{(k)}^\top\hat{\mathbf{z}}_{(k)} - \frac{2}{Ns}\hat{\mathbf{z}}^\top\hat{\mathbf{X}}\hat{\mathbf{X}}^\top\hat{\mathbf{z}} + \frac{1}{Ns}\hat{\mathbf{z}}^\top\hat{\mathbf{X}}\hat{\mathbf{X}}^\top\hat{\mathbf{z}} && (6.41) \\
&= \hat{\mathbf{z}}^\top(\frac{1}{s}\oplus\sum_{k=1}^{K}\frac{1}{n_k}\hat{\mathbf{X}}_{(k)}\hat{\mathbf{X}}_{(k)}^\top)\hat{\mathbf{z}} - \frac{1}{Ns}\hat{\mathbf{z}}^\top\hat{\mathbf{X}}\hat{\mathbf{X}}^\top\hat{\mathbf{z}} && (6.42) \\
&= \hat{\mathbf{z}}^\top(C-A)\hat{\mathbf{z}} && (6.43)
\end{aligned}
$$

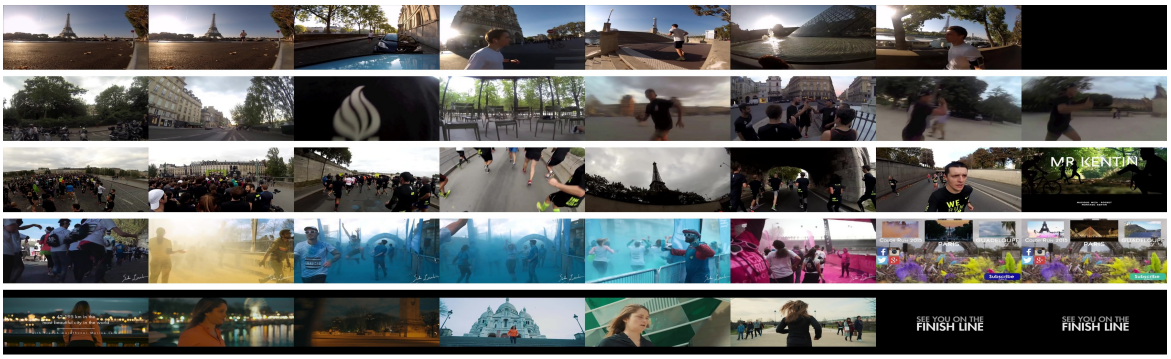
(6.40) is derived $\sum_{k=1}^{K} n_k = N$. $\boldsymbol{\mu}_k = \frac{1}{n_k s}\hat{\mathbf{X}}_{(k)}^\top\hat{\mathbf{z}}_{(k)}$ and $\bar{\boldsymbol{\mu}} = \frac{1}{Ns}\hat{\mathbf{X}}^\top\hat{\mathbf{z}}$ are used for (6.41).

6.12 Examples of dataset

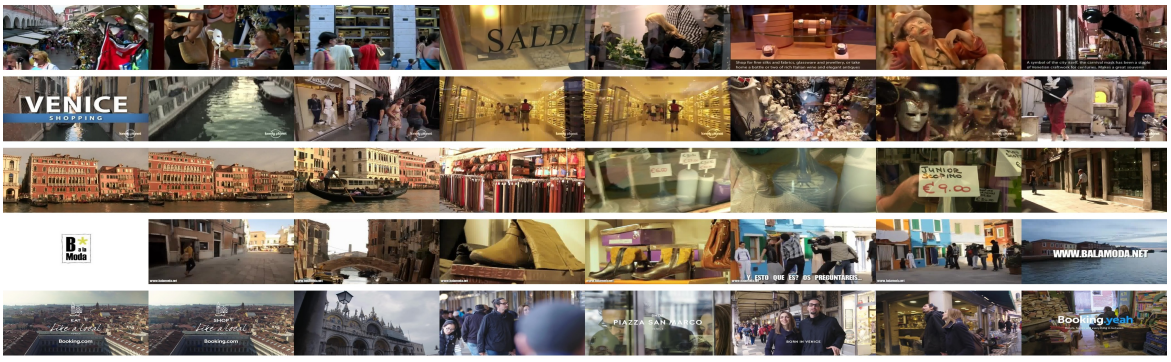
We show randomly selected frames of videos of our dataset in the following figures. The order of figure corresponds to the ones written in the Table. 1 in the main paper, namely, in the order of **TG**, **RG1**, **RG2**, and from top-row to bottom-row. Each row of figures corresponds to one video.



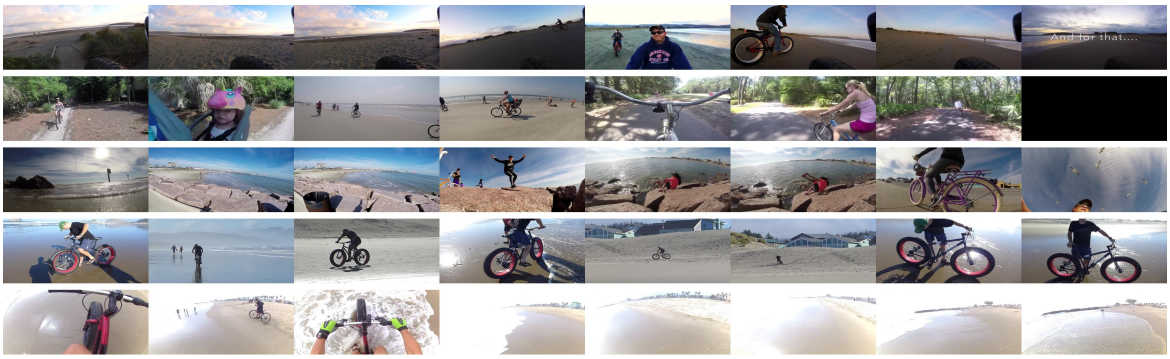
(a) Randomly selected frames from videos belonging to class run venice (**TG**). Each row corresponds to one video.



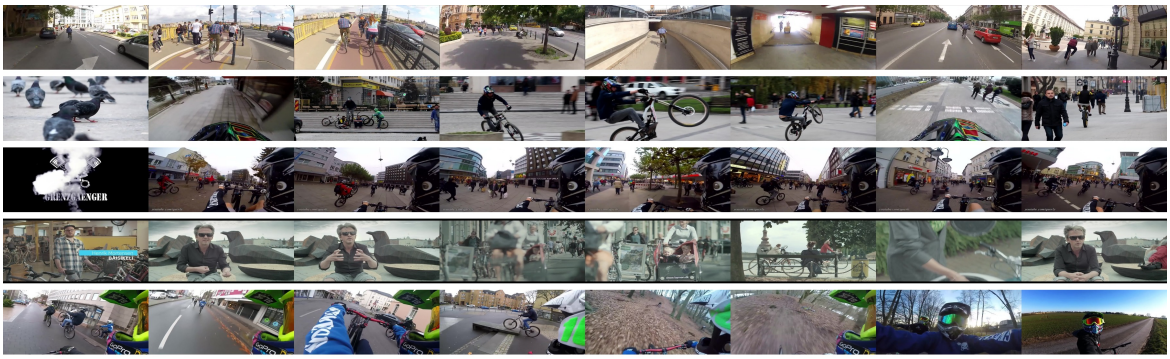
(b) Randomly selected frames from videos belonging to class run paris (**OG1**). Each row corresponds to one video.



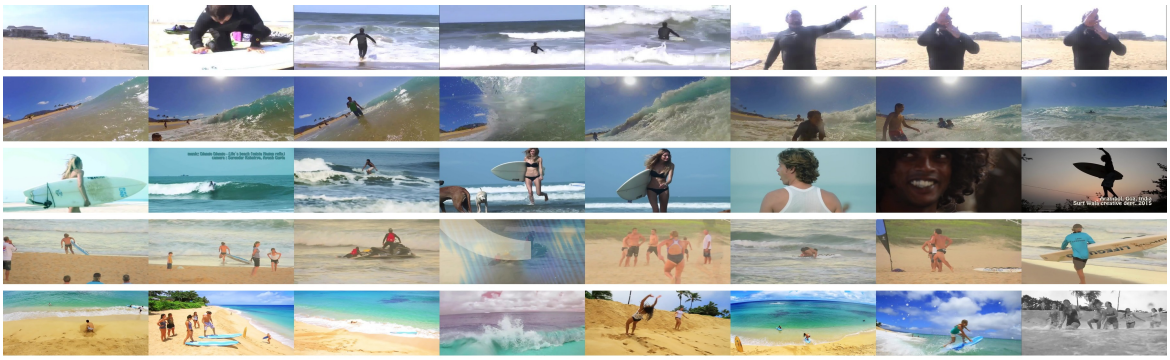
(c) Randomly selected frames from videos belonging to class shopping venice (**OG2**). Each row corresponds to one video.



(a) Randomly selected frames from videos belonging to class ride bike beach (**TG**). Each row corresponds to one video.

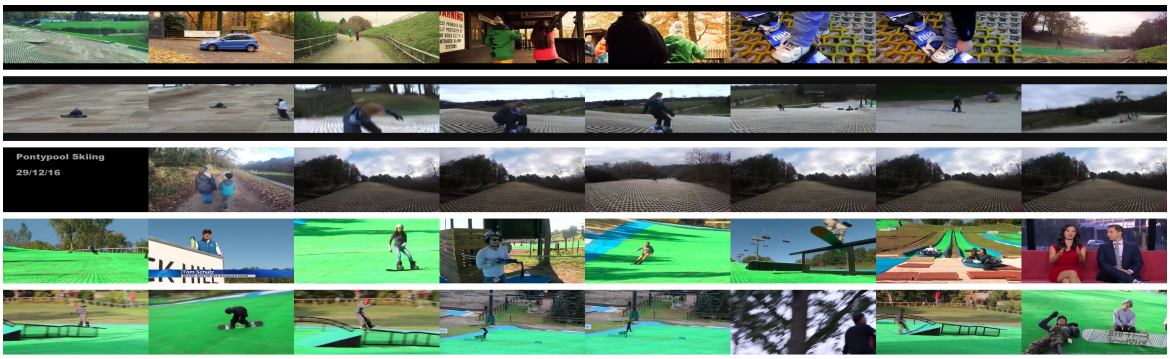


(b) Randomly selected frames from videos belonging to class ride bike city (**OG1**). Each row corresponds to one video.



(c) Randomly selected frames from videos belonging to class surf beach (**OG2**). Each row corresponds to one video.

(a) Randomly selected frames from videos belonging to class boarding snow mountain (**TG**). Each row corresponds to one video.



(b) Randomly selected frames from videos belonging to class boarding dry sloop (**OG1**). Each row corresponds to one video.

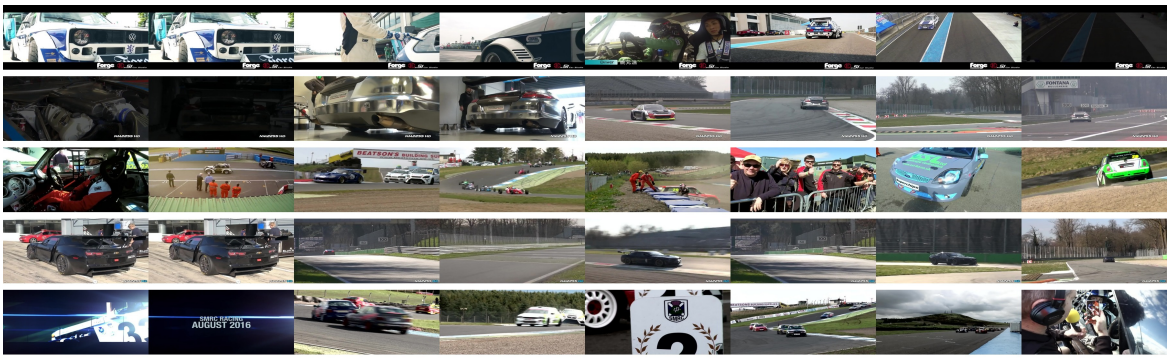
(c) Randomly selected frames from videos belonging to class hiking snow mountain (**OG2**). Each row corresponds to one video.

(a) Randomly selected frames from videos belonging to class dog chase sheep (**TG**). Each row corresponds to one video.

(b) Randomly selected frames from videos belonging to class dog play with kids (**OG1**). Each row corresponds to one video.

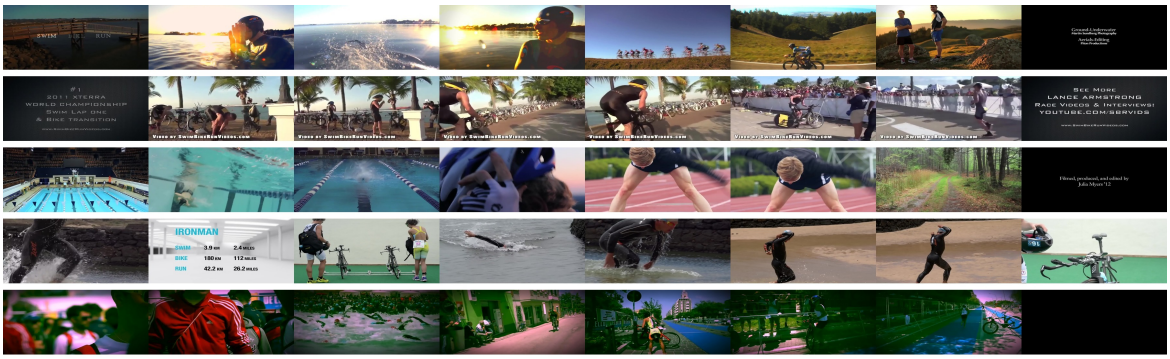
(c) Randomly selected frames from videos belonging to class sheep graze grass (**OG2**). Each row corresponds to one video.

(a) Randomly selected frames from videos belonging to class racing desert (**TG**). Each row corresponds to one video.

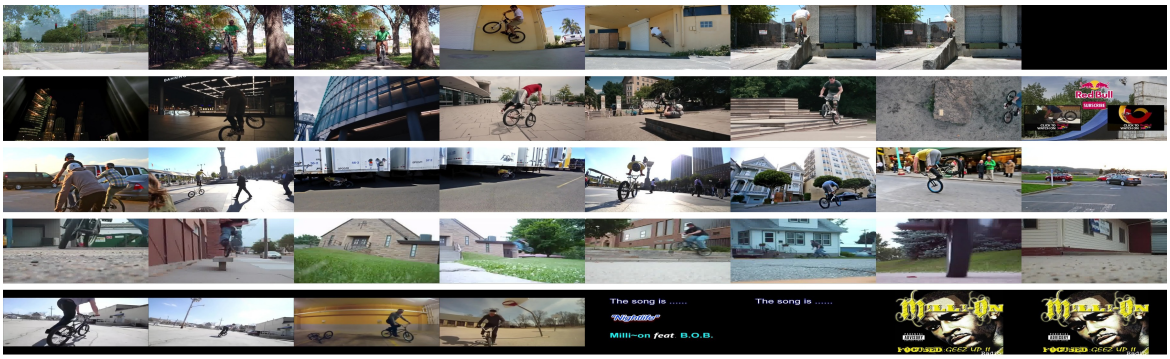


(b) Randomly selected frames from videos belonging to class racing circuit (**OG1**). Each row corresponds to one video.

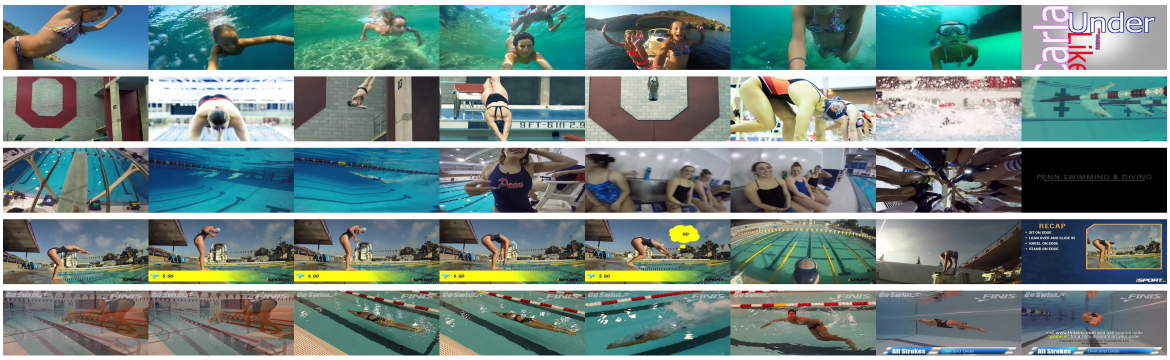
(c) Randomly selected frames from videos belonging to class riding camel desert (**OG2**). Each row corresponds to one video.



(a) Randomly selected frames from videos belonging to class swim riding bike (**TG**). Each row corresponds to one video.



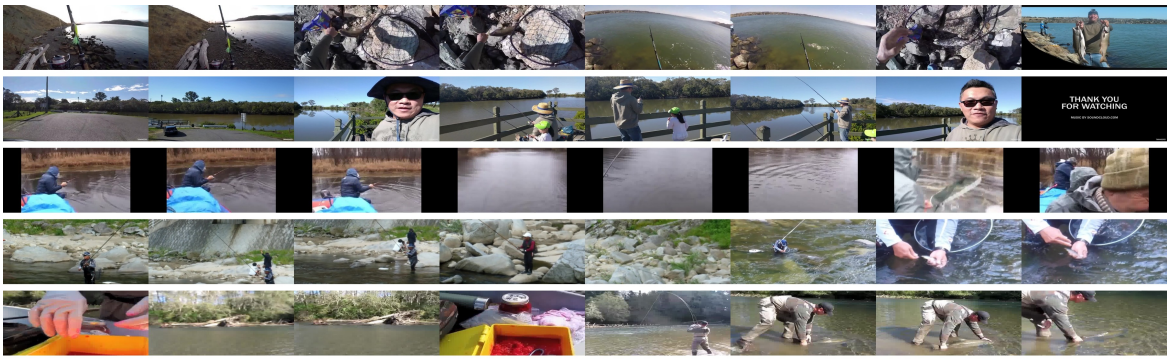
(b) Randomly selected frames from videos belonging to class riding bike trick (**OG1**). Each row corresponds to one video.



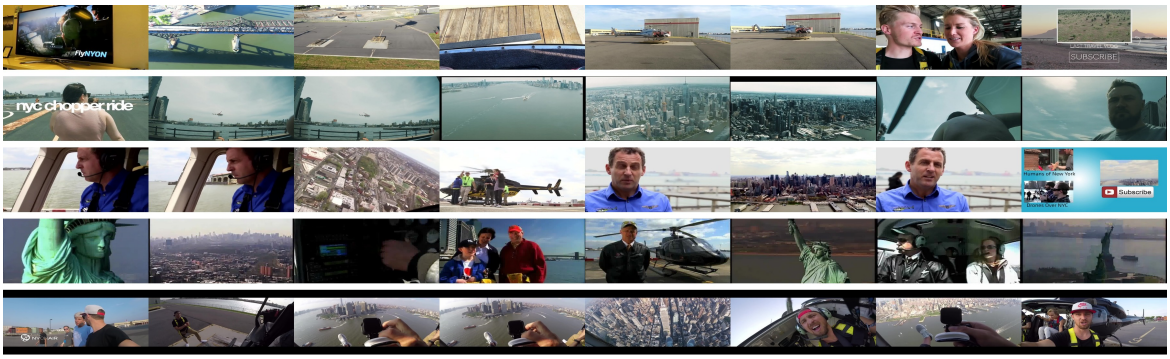
(c) Randomly selected frames from videos belonging to class swim dive (**OG2**). Each row corresponds to one video.

(a) Randomly selected frames from videos belonging to class fishing cook fish (**TG**). Each row corresponds to one video.

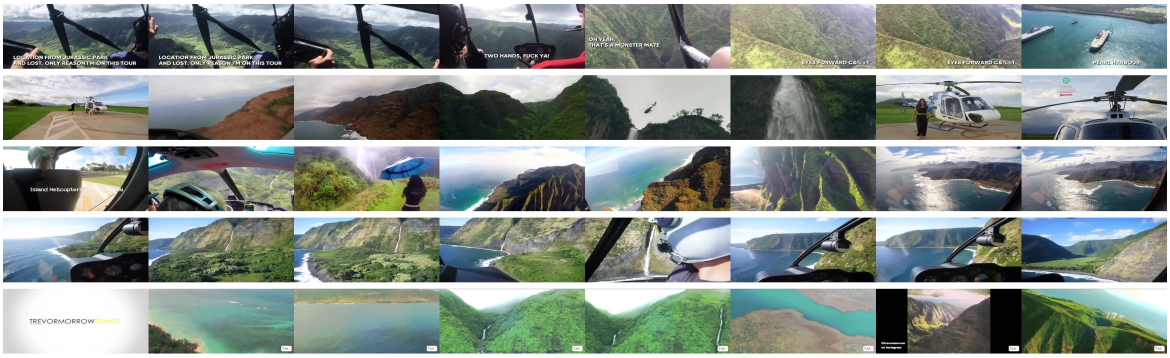
(b) Randomly selected frames from videos belonging to class cook fish village (**OG1**). Each row corresponds to one video.



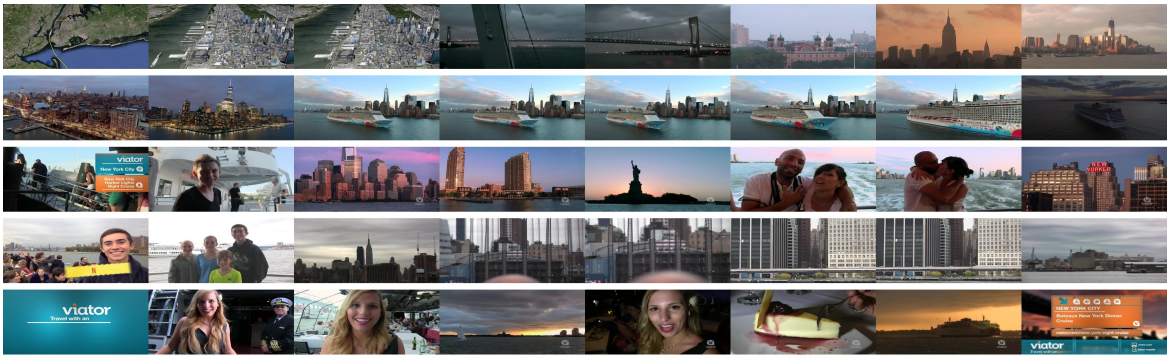
(c) Randomly selected frames from videos belonging to class fishing river (**OG2**). Each row corresponds to one video.



(a) Randomly selected frames from videos belonging to class helicopter NewYork (**TG**). Each row corresponds to one video.



(b) Randomly selected frames from videos belonging to class helicopter Hawaii (**OG1**). Each row corresponds to one video.



(c) Randomly selected frames from videos belonging to class NewYork cruse (**OG2**). Each row corresponds to one video.

(a) Randomly selected frames from videos belonging to class slackline rock climbing (**TG**). Each row corresponds to one video.

(b) Randomly selected frames from videos belonging to class rock climbing camping (**OG1**). Each row corresponds to one video.

(c) Randomly selected frames from videos belonging to class slcakline jaggling (**OG2**). Each row corresponds to one video.

(a) Randomly selected frames from videos belonging to class ride horse safari (**TG**). Each row corresponds to one video.

(b) Randomly selected frames from videos belonging to class ride horse mountain (**OG1**). Each row corresponds to one video.

(c) Randomly selected frames from videos belonging to class ride vehicle safari (**OG2**). Each row corresponds to one video.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

The goal of this thesis is to build a system that can generate explanations with multimodal information for categorization of a visual instance, and we propose a general framework to achieve it.

The desired explanation is the one which satisfies following two properties as:

(a) output explanations should be interpretable for humans,

(b) output explanations should be discriminative to the target category,

(c) the system should be applicable to modals regardless with/without supervised information,

(d) output explanations of different modals should be complemental to each other.

We proposed a novel framework for generating multimodal explanations based on the information theory, especially on the maximization of the interaction information, which is an extension of the mutual information defined on more than two variables.

In this thesis, we especially focus on the classification problem as the target of the explanation, where a visual instance $\mathbf{x}$ (e.g., image or video) is categorized to a class $\mathbf{y}$ by the process $\mathbf{f} : \mathbf{x} \rightarrow \mathbf{y}$. Under this premise, the task of explanation is divided by the two axes as follows:

- The difference of the decision process $\mathbf{f}$. One is when humans perceive visual information and categorize it. The other is by machines. We especially deal with the complicated machine learning models such as deep CNNs utilized as the de-facto standard in the current visual recognition task.

- The difference of dealing with positiveness/negativeness of the label **y**. In other words, the difference of explaining "why **x** is categorized to **y**" or "why **x** is **not** categorized to **y**."

Explanations for the Positiveness of Machines' Prediction (Chapter 4)

We considered the complementarity of multimodal explanations. We specifically treated the combination of linguistic and set of examples, where assigning supervised information is impossible. We discuss what the complementarity actually is from the viewpoint of interaction information on this task, and claimed that a complemental set to a linguistic explanation is a "*discriminative set*" of examples by which not only category label but also the linguistic explanation is identifiable from it. We proposed to parameterize the joint probability of variables to explain, and to be explained by the three neural networks. To explicitly treat the complementarity, auxiliary models responsible for the explanations were trained simultaneously to maximize the approximated lower bound of the interaction information. We empirically demonstrated the effectiveness of the method by the experiments conducted on the two visual recognition datasets.

Explanations for the Negativeness of Machines' Prediction (Chapter 5)

We particularly focused on the explanation for the negativeness of the prediction. In other words, we attempted to build a model that not only categorizes a sample but also generates multi-modal explanations the reason for "why X is not predicted **not** A but B," referring this type of explanations as *counter-factual* explanations. Especially, we treated a video as the target instance dealing with a spatiotemporal region of the target video and (the existence of) an attribute as elements.

The expected output of the visual-linguistic explanation should have the following two properties: (1) Visual explanation is the region which retains high positiveness/negativeness on model prediction for specific positive/negative classes, (2) Linguistic explanation is compatible with the visual counterpart. The score to measure how the requirements above are fulfilled is referred to as the *counterfactuality*, and we proposed a novel algorithm to predict *counterfactuality* while identifying the important region for the linguistic explanation. The proposed algorithm can be seen as the maximization of the lower-bound of the proposed framework. We demonstrated the effectiveness of the approach on two existing datasets extended in this work.

Explanations for the Positiveness of Humans' Prediction (Chapter 6)

We investigated generating multimodal explanations, which is the combination of linguistic and example-based explanations, for the positiveness of humans' prediction.

As a use case when the explanation is required, we considered the situation that the subjective label is provided by a human. Particularly, we set our goal to generate example-based explanations by video summarization when we have several groups (such as by preference) of videos and we cannot know why they are divided in such way they are.

To obtain a summarization containing interpretable and important parts, we assumed the properties the desired output should be are (1) diverse, (2) representative, and (3) discriminative. To satisfy (1)-(3) simultaneously, we proposed a novel criterion for video summarization, that is, the weighted sum of inner-video, inner-class, between-class variances inspired by Fisher criterion defined on the features of each segment of videos. To solve it efficiently, We also proposed a novel algorithm based on Concave-Convex Procedure (CCCP), which can be simply solved and the convergence is guaranteed.

This practical algorithm can be obtained from our framework based on interaction information when assuming that all segments are sampled hierarchically from the Gaussian distribution.

We compiled a novel dataset for evaluating this task, we demonstrated the effectiveness of proposed method under the limitation of the discriminative linguistic information can be obviously detected from the category labels, by performing the qualitative and quantitative experiments on it.

Bibliography

[com] Oxford living dictionaries. Oxford University Press. complement https://en.oxforddictionaries.com/definition/complement.

[2] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

[3] Anne Hendricks, L., Hu, R., Darrell, T., and Akata, Z. (2018). Grounding visual explanations. In *ECCV*.

[4] AUEB, M. T. R. (2016). One-vs-each approximation to softmax for scalable estimation of probabilities. In *Advances in Neural Information Processing Systems*, pages 4161–4169.

[5] Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., and Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

[6] Bentley, J. (1984). Programming pearls: algorithm design techniques. *Communications of the ACM*, 27(9):865–873.

[7] Chen, F. and De Vleeschouwer, C. (2011). Formulating team-sport video summarization as a resource allocation problem. *TCSVT*, 21(2):193–205.

[8] Chen, J., Song, L., Wainwright, M. J., and Jordan, M. I. (2018). Learning to explain: An information-theoretic perspective on model interpretation. In *ICML*.

[9] Chollet, F. (2015). keras. https://github.com/fchollet/keras.

[10] Chu, W.-S., Song, Y., and Jaimes, A. (2015). Video co-summarization: Video summarization by visual co-occurrence. In *CVPR*.

[11] Dabkowski, P. and Gal, Y. (2017). Real time image saliency for black box classifiers. In *Advances in Neural Information Processing Systems*, pages 6967–6976.

[12] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR*.

[13] Doersch, C., Gupta, A., and Efros, A. A. (2013). Mid-level visual element discovery as discriminative mode seeking. In *NIPS*.

[14] Doersch, C., Singh, S., Gupta, A., Sivic, J., and Efros, A. (2012). What makes paris look like paris? *ACM Transactions on Graphics*, 31(4).

[15] Doersch, C., Singh, S., Gupta, A., Sivic, J., and Efros, A. A. (2015). What makes paris look like paris? *Communications of the ACM*, 58(12).

[16] Elhamifar, E. and Kaluza, M. C. D. P. (2017). Online summarization via submodular and convex optimization. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[17] Elhamifar, E., Sapiro, G., and Vidal, R. (2012). See all by looking at a few: Sparse modeling for finding representative objects. In *CVPR*.

[18] Fleischman, M., Roy, B., and Roy, D. (2007). Temporal feature induction for baseball highlight classification. In *ACMMM*.

[19] Fong, R. C. and Vedaldi, A. (2017). Interpretable explanations of black boxes by meaningful perturbation. In *ICCV*.

[20] Gao, S., Ver Steeg, G., and Galstyan, A. (2016). Variational information maximization for feature selection. In *NIPS*.

[21] Gong, B., Chao, W.-L., Grauman, K., and Sha, F. (2014). Diverse sequential subset selection for supervised video summarization. In *NIPS*.

[22] Grant, M. and Boyd, S. (2008). Graph implementations for nonsmooth convex programs. In *Recent Advances in Learning and Control*. http://stanford.edu/~boyd/graph_dcp.html.

[23] Grant, M. and Boyd, S. (2014). CVX: Matlab software for disciplined convex programming, version 2.1. http://cvxr.com/cvx.

[24] Gu, C., Sun, C., Ross, D. A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., Schmid, C., and Malik, J. (2018). Ava: A video dataset of spatio-temporally localized atomic visual actions. In *CVPR*.

[25] Gygli, M., Grabner, H., Riemenschneider, H., and Van Gool, L. (2014). Creating summaries from user videos. In *ECCV*.

[26] Gygli, M., Grabner, H., and Van Gool, L. (2015). Video summarization by learning submodular mixtures of objectives. In *CVPR*.

[27] Hara, K., Kataoka, H., and Satoh, Y. (2018). Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6546–6555.

[28] He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[29] Hendricks, L. A., Akata, Z., Rohrbach, M., Donahue, J., Schiele, B., and Darrell, T. (2016). Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer.

[30] Hong, R., Tang, J., Tan, H.-K., Yan, S., Ngo, C., and Chua, T.-S. (2009). Event driven summarization for web videos. In *SIGMM workshop*.

[31] Jain, A., Gupta, A., Rodriguez, M., and Davis, L. S. (2013). Representing videos using mid-level discriminative patches. In *CVPR*.

[32] Jang, E., Gu, S., and Poole, B. (2016). Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

[33] Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*.

[34] Jiang, Y.-G., Liu, J., Roshan Zamir, A., Toderici, G., Laptev, I., Shah, M., and Sukthankar, R. (2014). THUMOS challenge: Action recognition with a large number of classes. http://crcv.ucf.edu/THUMOS14/.

[35] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., and Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *CVPR*.

[36] Khosla, A., Hamid, R., Lin, C.-J., and Sundaresan, N. (2013). Large-scale video summarization using web-image priors. In *CVPR*.

[37] Kim, G., Sigal, L., and Xing, E. P. (2014). Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In *CVPR*.

[38] Koh, P. W. and Liang, P. (2017). Understanding black-box predictions via influence functions. In *ICML*.

[39] Kong, S., Shen, X., Lin, Z., Mech, R., and Fowlkes, C. (2016). Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*.

[40] Kulesza, A., Taskar, B., et al. (2012). Determinantal point processes for machine learning. *Foundations and Trends® in Machine Learning*, 5(2–3):123–286.

[41] Lanckriet, G. R. and Sriperumbudur, B. K. (2009). On the convergence of the concave-convex procedure. In *NIPS*.

[42] Lee, Y. J., Ghosh, J., and Grauman, K. (2012). Discovering important people and objects for egocentric video summarization. In *CVPR*.

[43] Li, Y., Liu, L., Shen, C., and van den Hengel, A. (2015). Mid-level deep pattern mining. In *CVPR*.

[44] Liu, D., Hua, G., and Chen, T. (2010). A hierarchical visual model for video object summarization. *TPAMI*, 32(12):2178–2190.

[45] Liu, T. and Kender, J. R. (2002). Optimization algorithms for the selection of key frame sequences of variable length. In *ECCV*.

[46] Lu, Z. and Grauman, K. (2013). Story-driven summarization for egocentric video. In *CVPR*.

[47] Ma, Y.-F., Lu, L., Zhang, H.-J., and Li, M. (2002). A user attention model for video summarization. In *ACMMM*.

[48] Mac Aodha, O., Su, S., Chen, Y., Perona, P., and Yue, Y. (2018). Teaching categories to human learners with visual explanations. In *CVPR*.

[49] Maddison, C. J., Mnih, A., and Teh, Y. W. (2016). The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*.

[50] Mahasseni, B., Lam, M., and Todorovic, S. (2017). Unsupervised video summarization with adversarial lstm networks. In *CVPR*.

[51] McGill, W. (1954). Multivariate information transmission. *Transactions of the IRE Professional Group on Information Theory*, 4(4):93–111.

[52] Ngo, C.-W., Ma, Y.-F., and Zhang, H.-J. (2003). Automatic video summarization by graph modeling. In *ICCV*.

[53] Niebles, J. C., Chen, C.-W., and Fei-Fei, L. (2010). Modeling temporal structure of decomposable motion segments for activity classification. In *European conference on computer vision*, pages 392–405. Springer.

[54] Panda, R., Das, A., Wu, Z., Ernst, J., and Roy-Chowdhury, A. K. (2017). Weakly supervised summarization of web videos. In *ICCV*.

[55] Panda, R. and Roy-Chowdhury, A. K. (2017). Collaborative summarization of topic-related videos. In *CVPR*.

[56] Park, D. H., Hendricks, L. A., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., and Rohrbach, M. (2018). Multimodal explanations: Justifying decisions and pointing to the evidence. In *31st IEEE Conference on Computer Vision and Pattern Recognition*.

[57] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch.

[58] Plummer, B. A., Brown, M., and Lazebnik, S. (2017). Enhancing video summarization via vision-language embedding. In *Computer Vision and Pattern Recognition*.

[59] Potapov, D., Douze, M., Harchaoui, Z., and Schmid, C. (2014). Category-specific video summarization. In *ECCV*.

[60] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should i trust you?: Explaining the predictions of any classifier. In *SIGKDD*. ACM.

[61] Ross, A. S., Hughes, M. C., and Doshi-Velez, F. (2017). Right for the right reasons: Training differentiable models by constraining their explanations.

[62] Samangouei, P., Saeedi, A., Nakagawa, L., and Silberman, N. (2018). Explaingan: Model explanation via decision boundary crossing transformations. In *The European Conference on Computer Vision (ECCV)*.

[63] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*.

[64] Sharghi, A., Gong, B., and Shah, M. (2016). Query-focused extractive video summarization. In *ECCV*.

[65] Sharghi, A., Laurel, J. S., and Gong, B. (2017). Query-focused video summarization: Dataset, evaluation, and a memory network based approach. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2127–2136. IEEE.

[66] Simonyan, K., Vedaldi, A., and Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

[67] Singh, G., Saha, S., Sapienza, M., Torr, P., and Cuzzolin, F. (2017). Online real time multiple spatiotemporal action localisation and prediction.

[68] Singh, S., Gupta, A., and Efros, A. A. (2012). Unsupervised discovery of mid-level discriminative patches.

[69] Song, Y., Vallmitjana, J., Stent, A., and Jaimes, A. (2015). Tvsum: Summarizing web videos using titles. In *CVPR*.

[70] Soomro, K., Zamir, A. R., and Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

[71] Sun, M., Farhadi, A., and Seitz, S. (2014). Ranking domain-specific highlights by analyzing edited videos. In *ECCV*.

[72] Tokui, S., Oono, K., Hido, S., and Clayton, J. (2015). Chainer: a next-generation open source framework for deep learning. In *Proceedings of Workshop on Machine Learning Systems (LearningSys) in The Twenty-ninth Annual Conference on Neural Information Processing Systems (NIPS)*.

[73] Tolomei, G., Silvestri, F., Haines, A., and Lalmas, M. (2017). Interpretable predictions of tree-based ensembles via actionable feature tweaking. In *KDD*.

[74] Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

[75] Tran, D., Ray, J., Shou, Z., Chang, S.-F., and Paluri, M. (2017). Convnet architecture search for spatiotemporal feature learning. *arXiv preprint arXiv:1708.05038*.

[76] Tran, D., Yuan, J., and Forsyth, D. (2014). Video event detection: From subvolume localization to spatio-temporal path search.

[77] Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. In *NIPS*.

[78] Wachter, S., Mittelstadt, B., and Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the gdpr.

[79] Wah, C., Branson, S., Welinder, P., Perona, P., and Belongie, S. (2011). The Caltech-UCSD Birds-200-2011 Dataset. Technical report.

[80] Ye, J., Zhao, Z., and Wu, M. (2008). Discriminative k-means for clustering. In *Advances in neural information processing systems*, pages 1649–1656.

[81] Yuille, A. L. and Rangarajan, A. (2002). The concave-convex procedure (cccp). In *NIPS*.

[82] Yuille, A. L. and Rangarajan, A. (2003). The concave-convex procedure. *Neural computation*, 15(4):915–936.

[83] Zhang, J., Lin, Z., Brandt, J., Shen, X., and Sclaroff, S. (2016a). Top-down neural attention by excitation backprop. In *European Conference on Computer Vision*, pages 543–559. Springer.

[84] Zhang, K., Chao, W.-L., Sha, F., and Grauman, K. (2016b). Summary transfer: Exemplar-based subset selection for video summarization. In *CVPR*.

[85] Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929.

[86] Zhou, B., Sun, Y., Bau, D., and Torralba, A. (2018). Interpretable basis decomposition for visual explanation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 119–134.

[87] Zhu, G., Huang, Q., Xu, C., Rui, Y., Jiang, S., Gao, W., and Yao, H. (2007). Trajectory based event tactics analysis in broadcast sports video. In *ACMMM*.

Publications

Reviewed Conference

1. Atsushi Kanehira, Luc Van Gool, Yoshitaka Ushiku, and Tatsuya Harada, "*Viewpoint*-aware Video Summarization," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2018), Saltlake City, Jun., 2018.

2. Atsushi Kanehira and Tatsuya Harada, "Multi-label Ranking from Positive and Unlabeled Data," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, Jun., 2016.

3. Katsunori Ohnishi, Atsushi Kanehira, Asako Kanezaki, and Tatsuya Harada, "Recognizing Activities of Daily Living with a Wrist-mounted Camera," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016), Las Vegas, Jun., 2016.

4. Atsushi Kanehira, Andrew Shin, and Tatsuya Harada, "True-negative Label Selection for Large-scale Multi-label Learning," Proceedings of IEEE International Conference on Pattern Recognition (ICPR 2016), Cancun, Dec., 2016.

Others

1. **(Invited talk)** Atsushi Kanehira and Tatsuya Harada, "Multi-label Ranking from Positive and Unlabeled Data (CVPR 2016)," the 19th Meeting on Image Recognition and Understanding, Aug., 2016.

2. **(Invited talk)** Katsunori Ohnishi, Atsushi Kanehira, Asako Kanezaki, and Tatsuya Harada, "Recognizing Activities of Daily Living with a Wrist-mounted Camera (CVPR 2016)," the 19th Meeting on Image Recognition and Understanding, Aug., 2016.

3. **(Competition, invited oral)** Senthil Purushwalkam, Yuichiro Tsuchiya, Atsushi Kanehira, Asako Kanezaki, and Tatsuya Harada, the 4th place in the task 2a: Classification+localization with provided training data, ImageNet Large Scale Visual Recognition Challenge 2014 in conjunction with ECCV 2014.