Automatic Annotation of Online Multimedia Data with Similarity Relations

Yixin Chen

Doctor of Philosophy

School of Computer Science

McGill University

Montreal,Quebec

2017-12-15

A thesis submitted to McGill University in partial fulfillment of the requirements of the degree of Doctor of Philosophy

ABSTRACT

Fueled by the advancements in multimedia and networking technologies, recent years have witnessed the explosive growth and proliferation of online multimedia data hosting and sharing services (e.g., Facebook, YouTube, Instagram, Snapchat). Online multimedia data (e.g., photos, video clips) have become the biggest big data. Compared with other types of data, they are opaque to machines, thus less manageable, searchable, or reusable. Therefore, social networking platforms such as Flickr attach keywords or tags that describe the visual content to corresponding multimedia data for further management or retrieval tasks. However, the correspondence between the low-level visual features and high-level semantic meanings of multimedia data is complicated. Also, constructing high-quality training sets for learning this correspondence is challenging. In this thesis, based on Conditional Random Field, we propose an automatic approach to estimate the relevance of the visual content of multimedia data with candidate keywords or tags. This estimation is the foundation of a variety of real-world applications, for example, automatic annotation, multimedia data retrieval, spamming or polluted tags detection.

First, we focus the problem of intra-modal similarity or near-duplicate detection on the visual modality of online videos. To both fast and accurately conduct the detection, we represent a video with a bag of relatively simple feature vectors instead of a composite feature vector and propose a more parallelable feature extraction algorithm. By defining the concept of informativeness, we prove that the fusion of multiple feature vectors preserves more information about videos; thus it is more discriminative in the detection/retrieval task than the composite feature vectors.

With the system developed for the detection of near-duplicate videos,

we survey the textual modality (i.e., titles, tags, description) of a real-world web video dataset. Besides the statistical properties, we empirically verify the homophily assumption, which refers to the tendency that similar items bond together, on this dataset. Particularly, we validate the assumption from two perspectives: visually similar multimedia items are associated with semantically similar descriptions and vice versa.

Finally, based on the homophily assumption, we devise a mid-level multi-view multi-label relevance estimation approach to assess the cross-modal relevance between visual content and potential keywords. In this thesis, a view refers to the similarity relations between the multimedia items. Under each visual feature, the similarity relations between the items can be different. The differences and consistency between the views of multimedia data are incorporated into the approach to enhance the accuracy of estimation.

ABRÉGÉ

Alimentes par les progrs des technologies multimdia et de mise en rseau, ces dernires annes ont vu la croissance explosive et la prolifration des services d'hbergement et de partage de donnes multimdia en ligne (par exemple, Facebook, YouTube, Instagram, Snapchat). Les donnes multimdia en ligne (par exemple, les photos, les clips vido) sont devenues les plus grosses donnes. Compars d'autres types de donnes, ils sont opaques pour les machines, donc moins faciles grer, interrogeables ou rutilisables. Par consquent, les plates-formes de rseautage social telles que Flickr associent des mots-cls ou des balises qui dcrivent le contenu visuel des donnes multimdia correspondantes pour d'autres tches de gestion ou de rcupration. Cependant, la correspondance entre les caractristiques visuelles de bas niveau et les significations smantiques de haut niveau des donnes multimdias est complique. De plus, la construction d'ensembles d'entranement de haute qualit pour l'apprentissage de cette correspondance est difficile. Dans cette thse, base sur un champ alatoire conditionnel, nous proposons une approche automatique pour estimer la pertinence du contenu visuel de donnes multimdia avec des mots-cls ou des tags candidats. Cette estimation est le fondement d'une varit d'applications du monde rel, par exemple, l'annotation automatique, la rcupration de donnes multimdia, le spam ou la dtection de tags pollus.

indent Tout d'abord, nous focalisons le problme de la similarit intra-modale ou de la dtection quasi-duplique sur la modalit visuelle des vidos en ligne. Pour mener la fois rapidement et avec prcision la dtection, nous reprsentons une vido avec un sac de vecteurs caractristiques relativement simples au lieu d'un vecteur de caractristiques composites et proposons un algorithme d'extraction de caractristiques plus parallle. En dfinissant le concept d'informativit, nous

prouvons que la fusion de plusieurs vecteurs de caractristiques prserve plus d'informations sur les vidos; il est donc plus discriminant dans la tche de dtection / rcupration que les vecteurs de caractristiques composites.

indent Avec le systme dvelopp pour la dtection de vidos quasi-dupliques, nous tudions la modalit textuelle (c'est--dire, les titres, les tiquettes, la description) d'un ensemble de donnes vido web du monde rel. Outre les proprits statistiques, nous vrifions empiriquement l'hypothse d'homophilie, qui fait rfrence la tendance la liaison entre des lments similaires, sur cet ensemble de donnes. En particulier, nous validons l'hypothse partir de deux perspectives: des lments multimdia visuellement similaires sont associs des descriptions smantiquement similaires et vice versa. indent Enfin, sur la base de l'hypothse de l'homophilie, nous concevons une approche d'estimation de la pertinence multi-tiquettes multi-vues de niveau intermdiaire pour valuer la pertinence trans-modale entre le contenu visuel et les mots-cls potentiels. Dans cette thse, une vue fait rfrence aux relations de similarit entre les lments multimdias. Sous chaque caractristique visuelle, les relations de similarit entre les lments peuvent tre diffrentes. Les diffrences et la cohrence entre les vues de donnes multimdias sont incorpores dans l'approche pour amliorer la prcision de l'estimation.

ACKNOWLEDGEMENTS

First and foremost, I would like to offer my sincerest gratitude to my supervisor, Prof. Wenbo He, for the inspiration of ideas, guidance on my research, and the encouragement to help me through numerous setbacks and frustrations. I am fortunate to have you as my supervisor.

To my mother, Anhua Cai, without your unconditional love and support, I could not complete my study at McGill University.

Thank my colleagues and the members of my lab, Xinye Lin and Wen Wang. I enjoy the eye-opening discussions that we had about the research ideas.

Lastly, I would like to thank the National University of Defense Technology and the Chinese Scholarship Council, for their financial support of my study in McGill.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
Introduction

The focus of the work in this thesis is developing a method for the automatic annotation of web multimedia content (e.g., photos, video clips). Specifically, we intend to assign keywords or tags that correctly interpret the visual content to an image or a video. A roadmap toward this objective consists of three steps: (1) devising a method which facilitates the discovery of the similar or near-duplicate images/videos of a given image/video; (2) investigating the relationships between the visual and textual modalities of a real-world multimedia dataset; (3) developing a method that quantitatively gauges the relevance between these modalities of multimedia data. The last step predicates on the first two steps because of the following reasons: first, the proposed cross-modal relevance estimation method is based on the intra-modal similarity relationships, which are detected in step 1; second, the validity of the homophily assumption [34] of the method in step 3 is empirically verified in step 2.

1.1 Motivation

Since the era of Web 2.0, the dominant form of online data has evolved beyond the restraints of text and hyperlinks. Multimedia data (e.g., images, videos) have been deeply involved in the communications of the Internet users. Additionally, statistics [92] shows that visual content is more engaging, influential, and illustrative in marketing or social media. For example, colored visuals increase people's willingness to read a piece of material by 80%; 51.9% of marketing professionals worldwide name video as the type of content with the best ROI (Return on Investment); visual content is more than 40X more

likely to get shared on social media than other types of contents[92]. It was the technical issues rather than marketing strategies that prevent the propagation of multimedia data on the Internet.

Fueled by the advancements of multimedia and networking technologies, recent years have witnessed the explosive growth and proliferation of the online multimedia contents hosting and sharing services (e.g., Facebook, YouTube, Instagram, Snapchat). For instance, by 2015, Instagram users had shared over 30 billion photos, which had grown at the speed of 70 million per day [80]. In 2016, the rate had risen to more than 95 million per day [62]. Videos have been exhibiting a similar trend. According to Cisco Systems, Internet videos accounted for 78% of all U.S. Internet traffic in 2014 and is expected to rise to 84% in 2018 [88]. In the world's most popular online video sharing and hosting system, YouTube, the number of users has exceeded a billion [136], and it has been estimated that there are over 300 hours of video clips uploaded per minute [136]. Multimedia data are increasingly important to social media networks. Four of five top-tier social media brands are related to photos, which are Facebook, Instagram, Pinterest, Snapchat, and Twitter [8]. In the era of big data, multimedia data has been called the "biggest big data" [50].

Converting light signals to ideas is "a complex task far beyond the abilities of the world's most powerful computers" [12]. In human bodies, it involves numerous photoreceptor cells, ganglion cells, neurons in optical nerve and visual cortex. Therefore, compared with other types of data, multimedia data is opaque to machines, thus less manageable, searchable, or reusable.

Due to the opaqueness and sheer volume, it becomes more difficult to automatically detect and curb the diffusion of visually redundant multimedia data on the Internet, which lessens the effectiveness of the retrieval and

management optimization measures. For example, removing or marking redundant images or videos can prevent the search engine from returning pages of duplicated items, thus allowing the users to more rapidly obtain the results that they expect.

The notion of redundancy in this context is more subjective than objective. The classifiers for the detection task are trained to learn how many editions of the visual content will make users regard the item as a different image or video. The items garnered by the redundancy detection methods manifest semantic similarities, so semantic concepts might emerge in this process. Therefore, the identification of similarities between images or videos lays out the foundation for automatic annotation.

Most of the online multimedia services and social networks support the annotation of the multimedia data because this functionality is useful for retrieving the data that satisfy the intent of users. The volume of this annotated multimedia data is appealing to the annotation algorithms that demand large training data, for instance, the deep neural networks. However, the quality of these user-provided annotations is not satisfactory as the training set. Kennedy et al. [74] showed that only 50% of the tags are visually relevant to image contents. It also has been reported [36] that in collecting image from the Internet to construct ImageNet datasets, the average accuracy of each synset is 26%. Moreover, as happened on social network platforms such as YouTube, these tags or descriptions might be polluted by spamming [10].

For the sake of capitalizing on the abundant tagged multimedia data from the social networks and avoiding the problems mentioned above, an automatic cross-modal relevance estimation method is in demand. The identification of irrelevant textual or visual content is beneficial to improving the performance

of information retrieval and content recommendation. Meanwhile, the candidate images or videos verified by this method can be used in building a more extensive training set for deep neural networks. Furthermore, this estimation is an essential step towards the automatic comprehension of multimedia data.

1.2 Contributions

As mentioned above, the work of this thesis is comprised of three interconnected projects, which corresponds to the three steps of our roadmap. Thus the contributions of the thesis can be summarized as follows:

- **Fast and accurate similarity relation detection of online videos**: the detection and retrieval of duplicate or near-duplicate images is a well-established field. Currently, academic communities are more engaged in the partial-duplicated image detection/retrieval problem. However, the problem of identifying redundancy becomes much more complicated for videos, despite the fact that videos are essentially a collection of images and the techniques in the two fields are transferable. The dilemma between the effectiveness and efficiency of the detection/retrieval methods is more challenging to deal with. On the one hand, accurate detection/retrieval requires complex and high-dimensional feature vectors to embody a video; on the other hand, the complexity and dimensionality of the feature vectors make the processing of videos time-consuming, which can be exhaustive considering the scale of online videos. In order to design a similar or near-duplicate video detection/retrieval system, we represent a video with a bag of relatively simple feature vectors instead of a composite feature vector and adapt the feature extraction algorithms to be more parallelable. By defining the concept of informativeness, we prove that the fusion of multiple feature vectors is more informative,

thus more discriminative in the detection/retrieval than the composite feature vectors.

- **Empirical study of the statistical properties of an online video dataset.** Allowing users to describe or comment on the content of images or videos is prevalent in web multimedia services or social networks. From these descriptions, keywords are extracted and associated with the corresponding multimedia items to facilitate the retrieval or content management. These keywords form the textual modality of the online multimedia data. With the near-duplicate video detection/retrieval system developed, we survey the textual modality of a real-world web video dataset. Besides the statistical properties, we empirically verify the homophily assumption, which refers to the tendency that similar items bond together, on this dataset. Specifically, we validate the assumption from two perspectives: visually similar multimedia items are associated with semantically similar descriptions and vice versa.
- **Multi-view and multi-label relevance estimation between the visual and textual modalities.** Having validated the homophily assumption on a real-world video dataset, we propose a cross-modal relevance estimation method, which depends on the similarity detection methods on both of the modalities. To enhance the accuracy of the method, we utilize multiple visual feature extraction and representation algorithms on the visual modality. Under each feature, the similarity relations between the items vary thus form a distinct view. Based on a graphical learning model, Conditional Random Field, we design a mid-level multi-view relevance estimation approach. On the one hand, the fusion of multiple views is not conducted at the feature-level, where the objective is to construct a latent common subspace underlying these

views. Each of the views remains independent in our approach so that various visual features, including the ones produced by deep learning networks, can be applied directly. On the other hand, the fusion is not postponed to the point where the labels (i.e., keywords or tags) have been assigned to each multimedia item, so the diversity of each view can be more preserved in this way. Additionally, the graphical approaches are more suitable for the circumstances where an instance (i.e., image/video) can be described with more than one labels.

1.3 Outline

In Chapter 2, we briefly review the related work in the fields of near-duplicate detection/retrieval of multimedia data, the collaborative tagging systems or folksonomies that are deployed for the management of the online multimedia content, and the cross-modal relevance estimation methods. In Chapter 3, we introduce the design of our near-duplicate video detection/retrieval system, CompoundEyes. The focus is shifted to the statistical properties of the visual and textual modalities of the online video data in Chapter 4. Following the design of CompoundEyes and the statistical properties of online video datasets, we propose a multi-view and multi-label method to assess the relevance of the visual and textual modalities of the online multimedia data. In Chapter 6, we summarize our contribution and discuss future work.

CHAPTER 2
Related Work

2.1 Near-duplicate Multimedia Detection/Retrieval

In this section, we survey the approaches and techniques that have been applied in multimedia duplicate detection/retrieval. These methods can be divided into two categories: representing a multimedia item (i.e., an image or a video clip) as a processable data type (e.g., vectors), and enhancing the speed efficiency.

2.1.1 Feature Representation

Representing a multimedia item as a machine-processable data type is the foundation of multimedia duplicate detection/retrieval. Various methods have been developed in the literature. Despite the difference between the two multimedia objects (i.e., image and video), the representation techniques are transferable, since a video is composed of frames, which are images in essence.

Converting an image into a computer processable data type consists of extracting low-level features from the visual content, and describing the features with data types such as vectors. Generally speaking, based on the granularity, the features can be categorized into global or local features; so are the feature representations.

As implied by the name, global features capture the global properties of the image. Such examples include color distribution [48], color moment [140], texture [5], DCT (Discrete Cosine Transform) coefficients [112], etc. In contrast, local features, also known as local interest regions, are localized and salient regions of the image. These regions are covariant with the transformations (e.g., affine transformation) on the original image. A variety of region

detectors have been developed, which is summarized by Mikolajczyk et al. [93] Among these detectors, Lowe's Difference of Gaussian (DoG) detector [89] is frequently used.

Evidently, global features can only be described with global representations (e.g., fingerprints, signatures). For example, the color distribution of an image can be naturally represented as a multi-dimensional histogram [125]. The distance (e.g., L1, L2 norm, cosine distance) between the global representations of images indicates whether the two images are near-duplicate or not. This type of approaches are rapid and straightforward, but not robust to light changes, viewpoint changes, scale changes, occlusions, and so forth.

These drawbacks can be overcome by utilizing local feature-based approaches because local features are covariant with the transformations on the image thus capable of differentiating true near-duplicate pictures from false ones. There are two ways of representing local features. First, each local interest region of an image can be described with the information of the region, by using descriptors such as SIFT [89] or PCA-SIFT [73]. In this way, hundreds or thousands of local feature descriptors are generated for an image. With these descriptors, the problem of duplicate detection/retrieval can be solved by employing set matching [125]. However, since the matching between descriptors is not exclusive, other visual contextual information, such as spatial coherency [145, 129, 26], or geometric constraints [129, 86], are included to eliminate false matchings of images. In partial-duplicate image matching, this additional information is crucial to matching accuracy. Despite the effectiveness gain, the computational cost of these approaches is expensive, and this efficiency issue becomes worse in the context of extensive image data.

Second, the information of local interest regions of an image can also be summarized into a global representation. One representative method is known

as the BoWs (Bag-of-Words) method [108]. In this manner, the descriptors of all local interest regions are garnered, then clustering such as K-Means is performed on this set. Each center of the clusters is designated as a visual word and indexed. When the construction of the visual word vocabulary completes, every local region descriptor can be quantized and assigned to the index of the closest cluster center. After the quantization, an image is represented as a high-dimensional vector, each entry of which corresponds to the frequency of appearances of a visual word in the image. The BoWs method is robust to the transformations that fail the global features and avoids the high complexity of the set matching methods. However, this global representation is ineffective to spatial transformations, and the visual words are not expressive as text words. Additionally, the high-dimensional vector is sparse because the count of local interest regions in an image is limited. When the size of the vocabulary grows, the performance of this method will be saturated, and the discriminative ability of visual words drops.

There are research endeavors to improve the effectiveness of the BoWs method further. Hu et al. [56] group visual words into visual phrases, and enforce spatial coherence to eliminate false matches and reduce the quantization errors. Zhang et al. [144] filter visual words by descriptiveness with PageRank-like algorithms and construct visual phrases by co-occurrence. This co-occurring contextual information is also applied by Wei et al. [120] to correct typos in visual words. Chu et al. [26] propose a novel Combined-Orientation-Position (COP) consistency to refine visual words. The enforcement of coherence improve the effectiveness of the BoWs method but incurs more overhead to its already high computational cost.

Compared with images, videos have an additional temporal dimension. The most straightforward approach is neglecting this temporal information

thoroughly, and taking the average of the global representations of all the frames [27, 125]. On the other end of the spectrum is the sequence matching techniques, in which a video is regarded as a sequence of frames. Chiu et al. [22] formulate the matching problem as a shortest-path problem in the matching graph. Law-To et al. [79] track and label the trajectories of local interest points in videos. Huang et al. [58] track the changes of the content of frames from that of keyframes, and measures the sequence similarity with weighted edit similarity (WES). Zhou et al. [148] propose a 3-D tensor model to describe the changes of local interest region descriptors. Chiu et al. [23] transform the subsequence matching problem into 2D Hough space projection of pairwise frame similarities between two subsequences. Chou et al. [24] use a dynamic programming-like algorithm to match the symbolized video sequences. The computational cost of these sequence matching techniques is massive thus not favorable for web videos.

2.1.2 Speed Efficiency

Speed efficiency is a paramount concern for any practical multimedia near-duplicate detection/retrieval system, especially when the volume of data grows explosively. In the literature, there are two commonly applied approaches to accelerate the processing speed: filtering, and indexing.

With the filtering approach, the detection/retrieval system appears as a hierarchical system. The upper layer rapidly processes images/videos with light-weight feature extraction and representation methods and leaves the sophisticated instances to the lower layer with more complicated approaches. Zhao et al. [145] filter candidate near-duplicate images by comparing the BoWs representations of them, before entering the stage of local interest regions matching. In [125], Wu et al. build a hierarchical system with the global color histogram representation and local interest region representation based,

sliding window approaches. To further increase the processing speed, they [126] utilize the contextual information (e.g., thumbnail images, tags, titles, time durations, views, comments). The hierarchical design is also employed to accelerate the speed of the systems that apply sequence matching methods. Chiu et al. [23] select candidate near-duplicate video files by estimating the time-decay hit frequency. Chou et al. [24] utilize spatial-temporal index patterns to filter out non-duplicate videos. In the hierarchical design, it is difficult to determine when to switch from light-weight methods to more complicated ones, to keep the balance between speed and accuracy.

Indexing structures are used to expedite the retrieval of near-duplicate images or videos, after them being projected into feature space. Hash table is one of the most popular indexing structures. Other examples include LIP-IS [146, 149], LSH (Locality Sensitive Hashing) [27, 147], or inverted indexing [104].

Moreover, there are attempts to parallelize the processing of detection/retrieval. For example, Xie et al. [129] leverage the computational ability of GPU to hasten the time-consuming matrix calculations of the Harris-Hessian local feature detector. Hefeeda et al. [52] developed a distributed matching engine to find the K-nearest neighbors of high-dimensional multimedia representations. These efforts are essential to deploy near-duplicate multimedia detection/retrieval system in the cloud.

2.1.3 Feature Fusion

The academic community has investigated various feature fusion strategies to overcome the limitations of global and local feature representations. The combined features contain more information than the original features; thus they are more discriminative in detecting near-duplicate images or videos.

Shang et al. [104] propose two spatial-temporal features, one is based on Conditional Entropy (CE), the other is based on Local Binary Pattern (LBP). The CE and LBP methods are utilized to capture the spatial information within frames, whereas the temporal information is preserved by applying the w-shingling method. The resultant feature representations are high-dimensional and sparse, even after compression.

In [109, 110], by making use of the information of manifold, Song et al. translate key-frames into binary hash codes. The affinity relations of videos in HSV color distribution and LBP texture feature spaces are preserved in the training of the hash functions. A similar approach called kernelized multiple feature hashing (KMFH) was proposed by Zou et al. [155] for near-duplicate image detection. They embed the features of an image into kernel spaces. The resultant hash codes are discriminative and compact. However, the training process of these approaches employs optimization techniques, and the overhead of matrix computations is high.

Alternative fusion strategies include multiple instance learning [19] and ensemble fusion [5]. We employ both of these strategies in the field of near-duplicate video detection/retrieval. Zhang et al. [142] applied the Multiple Instance Learning approach into content-based image retrieval. The main concern of this work is to increase the retrieval accuracy rather than speed; thus the performance of this system in large dataset stays unknown. Amir [5] developed a feature fusion pipeline for visual concept detection. They extract various feature representations from videos, conduct unimodal training on each one of them, and combine the decisions via an ensemble fusion approach. The training algorithm is Support Vector Machine (SVM), which is not favorable for large dataset.

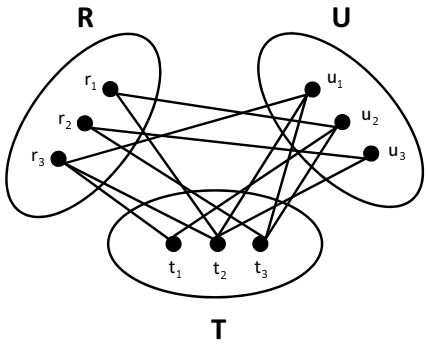


Figure 2–1: The Tripartite Graph Representation of Folksonomy

2.2 Folksonomy

In the Internet, images and videos are commonly accompanied with textual data, which not only illustrates the high-level semantic meanings of visual content to human beings but also is used by machines to locate videos in the database. The system that utilizes textual data to pinpoint resources has been known as folksonomy (i.e., collaborative tagging, social tagging).

Formally, folksonomy is defined as a set of 3-tuples $F = \{< r, t, u > | r \in R, t \in T, u \in U\}$, where R, T, and U are the collections of resources (i.e., visual content), textual data (e.g., tags), and users respectively. The relationships among these collections are commonly represented as a tripartite graph [87], as in Figure 2–1. From another perspective, this figure describes a process in which users add descriptions in the form of keywords to the shared content.

The annotations or tags of folksonomies reflect the collective "wisdom of the crowd". This semantically rich source of information is widely employed for the management of a variety of online services, such as Delicious or Flickr. The first and foremost concern of folksonomy is about the effectiveness of folksonomies. Through experiments on the social bookmarking site del.icio.us, Heymann et al. [54] concluded that the role that tags play in the websites could not be replaced by another source of information. By exploiting the structures of folksonomies, Hotho et al. [55] proposed an information retrieval model,

which ranks items based on the relevance to the search queries. Bragg et al. [11] designed a crowdsourcing taxonomy creation workflow that approaches the level of human experts, at the expense of more labor.

In the core of folksonomy construction is the learning of taxonomy from the tags or annotations of resources. This learning is challenging due to the imprecision and incompleteness of the tags. In [7], by comparing the performance of decentralized search that employs the hierarchical structures of learned folksonomies, Helic et al. evaluated four state-of-the-arts folksonomy algorithms, and concluded that DegCen/Cooc and CloCen/Cos algorithms are more superior. For the visual concept learning of image sharing websites such as Flickr, Zhu et al. [150] introduced semantic field and ontology-based semantic pooling to increase the relevancy of images to the target concepts, and the coverage of the concepts respectively. Aurnhammer et al. [7] proposed to incorporate visual features into the discovery of the relationships between data, for the retrieval of image sharing websites. The fusion of visual and semantic features is also employed in applications such as cross-modal retrieval, clustering, and recommendation.

One of the most prominent statistical properties of folksonomies is that the tag frequency distribution follows the power law, or Zipf's law distribution, which is an indicator of complex systems. Therefore, academic communities [40, 47, 17, 103, 78, 121] investigate the properties, structures, and dynamics of folksonomies as complex systems. The stabilization of tag distribution is also related to the emergence of vocabularies, or semantics, which is crucial to the effectiveness of folksonomies in organizing resources. Robu et al. [103] showed that the vocabularies extracted from folksonomies are considerably richer than those extracted from general search engine logs. Korner et al. [78] divided users into "categorizers" and "describers", and empirically examined

their impacts on the semantics of folksonomies. The results show that verbose taggers contribute more to the emergent semantics. Wetzker et al. [121] categorized folksonomies into narrow and broad folksonomies, by the annotation privilege. The former one restricts the tagging of resources to a limited number of users, whereas the latter one does not. They observe that individual tagging habits remain relatively stable, but the choice of tags still varies even for conceptually similar resources.

From another point of view, the appearance of a stable tag frequency distribution implies that the users of the folksonomy have reached a consensus about the vocabulary. Thus it is essential to ensure that this consensus can be achieved and to evaluate the effectiveness of the words in the vocabulary at isolating resources. Recent research concluded that the tagging choices tend to converge [21]. Besides, although according to the information theory, the effectiveness of tags in locating resources decreases as the growth of vocabulary, folksonomy systems work well in practice. It is believed that the semantic coherence between the tagging choices ensures this effectiveness.

The emergent semantics of tags in folksonomies is a high-level, rich source of information for the comprehension of the content of resources (e.g., multimedia data). The first step to exploit this form of information is to define a valid and pragmatic similarity measure. Cattuto et al. [16] evaluated five measures of tag relatedness: the co-occurrence count, three distributional measures that apply the cosine similarity, and a graph based measure. The semantic relations between the tags under these measures are compared against those of WordNet. The results indicate that the distributional measures establish paradigmatic relations between tags, and the combination of co-occurrence with popular tags achieve comparable performance to the most semantically accurate measures. Markines et al. [91] also grounded the collectively defined

semantics based on WordNet, and concluded that mutual information with distributional micro-aggregation yields the highest accuracy.

2.3 Cross-modal Relevance Estimation

From the era of Web 2.0, folksonomies (i.e., collaborative tagging systems) have been broadly applied in the retrieval, organization, and management of online multimedia data. The examples include Flickr, Picassa, YouTube. However, the tags or keywords contributed by users are too personalized thus lack coherence in semantics. To overcome these deficiencies of folksonomies, and sufficiently exploits the collective wisdom of users, academic communities attempt to automatically predict the tags that describe the visual content of corresponding multimedia data, which relies on the relevance estimation of the visual and textual modalities of multimedia items.

2.3.1 Cross-modal Learning and Multi-view Learning

Content-based automatic tagging is challenging due to the semantic gap in multimedia data, at the ends of which are the visual and textual modalities. The semantic gap can be illustrated as "the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation" [82]. Additionally, it has also been argued that automatically annotating multimedia data (e.g., images) in a general way is impossible, even if all the objects in the images or videos are detected and recognized. This argument is established on the assumption that multimedia items do not have intrinsic meaning, which only emerges from the interaction with users or another items [7].

Conventionally, the cross-modal relevance estimation problem is interpreted as a classification problem, where the features are extracted from the visual content, and the tags are treated as classes. In [35], Farhadi et al. predicted the attributes (i.e., classes) of the visual objects by feeding the visual

features of them into a linear SVM and logistic regression classifier. They noticed the effectiveness difference of visual features in predicting different attributes and selected the features accordingly in the classification. In [133], Yang et al. discovered that videos of the same class might have a substantial variation in visual content. They assumed the existence of latent classes and refined the tags with a bootstrapping ensemble scheme. To deal with the uncertainty and noise of the tags collected from users, Vahdat et al. [114] enabled the flip of training tags and penalized the number of flips in the training of SVM. This idea was also employed in the scenario of clustering [115].

In tasks such as clustering and retrieval, integrating the visual and textual features not only reduces semantically irrelevant images or videos but also is helpful for tag filtering or cleansing. Besides these the two modalities, the mid-level properties such as the social attributes of users [28] can also be exploited.

The integration of multiple modalities can be achieved by maximizing the correlated subspace (CCA), learning the common space or coding (Cross-modal Hashing), and extending the semantic topic models of words (e.g., LSA, pLSA, LDA), for instance. These approaches can be briefly described as follows:

1. The co-occurrence information of the visual and textual words can be used for estimating the correlations between the visual and textual features [123]. However, the CCA (Canonical Correlation Analysis) algorithm is more widely adopted because the visual and textual features belong to different spaces. The CCA-based algorithms can be kernelized (KCCA) to tackle the dimensionality issues [99, 44].
2. The cross-modal hashing methods belong to the manifold learning approaches, which can be used as both unsupervised [43] and supervised

[13, 151, 139] algorithms. The unsupervised algorithms are based on CCA, while the supervised algorithms exploit the semantic meanings of tags to train the hashing functions.

3. LSA (Latent Semantic Analysis), pLSA (probabilistic Latent Semantic Analysis), and LDA (Latent Dirichlet Allocation) are originally applied to extracting topics from textual documents. Academic communities have proposed various ways to extend them into the multi-modal scenarios. In [122], Wu et al. applied the LDA modalities to extract topics from time-sync video tags. Jia et al. [66] built an MRF (Markov Random Field) model over the LDA model, in which the similarities between multimedia items in the same modality or across different modalities are incorporated to boost the performance of cross-modal tasks. Lienhart et al. [83] developed a multi-layer pLSA model to merge the topics extracted from the visual and textual pLSA models.

Compared with the classification approaches, the CCA algorithms assume strong correlations between modalities, and the multi-modal LSA or LDA models restrain the relationships between modalities. On account of these limitations, and to cope with the large variations of the visual content of images or videos that belong to the same class, Zhuang et al. [154] extended the DL (dictionary learning) methods to the multi-modal settings. The method jointly learns dictionaries for each modality along with a cross-modal mapping function and discovered the shared structures between modalities.

The definitions of multi-view learning and multi-modal (or cross-modal) learning on multimedia data are quite similar. The meaning of view in this context does not refer to a viewpoint of a camera, but a perception of data under a feature. Multimedia data consists of modalities more than visual and textual modalities, for instance, RGB color distribution, the audio stream,

depth. Each of these modalities can be described by a feature or a set of features. Consequently, the techniques and methods that are adopted in these two class of learning algorithms are akin as well.

In a survey [130] of the multi-view learning algorithms, Xu et al. categorized the algorithms into three classes: co-training, multiple kernel learning, and sub-space based learning algorithms. The co-training algorithms are further compared in [94], and Christoudias et al. [25] devised a co-clustering style bootstrapping algorithm to resolve the disagreement between views. Multiple kernel learning algorithms are related to semi-supervised, especially graphical learning models. In [134], it is utilized to transfer the knowledge learned from images to the domain of videos. The sub-space based learning algorithms, including CCA [42], metric learning [116, 53], and latent space models [143], are also broadly used in cross-modal learning. Additionally, the multi-view learning algorithms can be extended into the domain of deep learning. Wang et al. [118] combined DNN-based approaches with linear and kernel CCA, and compared their performance.

2.3.2 Multi-label Learning

The semantic meanings of real-world images or videos are complicated. Therefore, assuming that each multimedia item belongs to only one semantic concept, which is fundamental for the supervised learning, is not sensible in this context. Multiple tags or labels can be associated with the visual content of an image or a video.

In [141], Zhang et al. reviewed representative algorithms for multi-label learning. These algorithms are divided into two categories; the first one transforms the multi-label learning problem into other well-studied learning problems, and the second one adapts different learning algorithms to fit the multi-label learning scenario. In the first category, the multi-label learning problem

can be transformed into multiple independent binary classification problems, a chain of binary classification problems, the label ranking problem, or an ensemble of multi-classification problems. In the second category, the candidate learning algorithms for the multi-label learning problems include kNN (k Nearest Neighbor), Decision Tree, maximum margin strategies such as SVM, optimization methods that are entropy-based. One of the most discernible differences between these multi-label learning algorithms is whether and how the correlations between labels are incorporated.

Among the learning techniques mentioned above, the maximum margin strategies (e.g., SVM) combined with graphical or kernel learning methods have attracted lots of attention from academic communities. In [105], Shen et al. employed multi-task SVM to model the inter-object relationships between objects in loosely-tagged images, while in [32], Fan et al. combined multi-task learning techniques with max-margin strategies to address the issues of inter-concept similarity and diversity. In these works, the task relatedness is utilized to embody the correlations between tags or labels.

The label propagation algorithm is a frequently used graphical algorithm for cross-modal, multi-view, and semi-supervised learning problems. It needs to be modified to account for the correlations between labels. In [68], Kang et al. co-propagated multiple labels simultaneously instead of one at a time. Kong et al. [77] propagated label sets rather than labels. Chen et al. [18] propagated labels on the ℓ_1 graph based on the Kullback-Leibler divergence which measures the similarity between labels. Wang et al. [119] developed another label set propagation algorithm that employs the social context features to capture the correlations between labels. Moreover, in the case of incomplete label assignment, which is typical for real-world application, a group

lasso technique is utilized to compensate for the errors in ranking the assigned labels against the unassigned labels [14].

Fusion techniques have also been applied to enhance the performance of multi-label learning algorithm. Wang et al. [117] fused the transition kernel and label kernel to incorporate label correlations and instance (i.e., multimedia items) similarities. Kong et al. [76] constructed a heterogeneous information network, which is capable of providing abundant information concerning the relationships among instances or label, for propagation. Xu et al. [131] constructed an affinity graph of labels based on the co-occurrence information, which is further fused with other visual affinity graphs induced from multiple features.

2.3.3 Semi-supervised Learning and Graphical Models

In the semi-supervised learning algorithms, the unlabeled multimedia instances are involved in the computations because of the assumption that the labeled and unlabeled data share common underlying similarity structures, both visual and semantical. Because the amount of unlabeled data is considerably more massive than that of the labeled data, the semi-supervised learning algorithms are preferable than supervised learning algorithms in plenty of scenarios. They are integrated with multi-view or multi-label learning techniques in the field of automatic annotation, such as Bayesian inference [15], co-training [132], or curriculum learning [41]. The purpose of the integration is the enhancement of accuracy, whereas the speed issues have rarely been the focus, except for the research conducted by Ravi et al. [101], where the streaming and distributed computing techniques have been applied to in the label propagation.

Graphical models are a subcategory of the semi-supervised learning algorithms. They are capable of naturally representing various relationships between or among multimedia items. Representative examples include the label propagation algorithm [152, 84, 106, 81, 59, 135, 38, 69, 138, 102, 127, 37, 9, 31] and the random-field-based models (e.g., Gaussian random field [153], conditional random field [63]).

A simple type of label propagation algorithms is neighbor voting [81, 135], in which the label of an instance is determined by the distribution of the labels of the instance's neighbors. When the neighbors of the neighbors are taken into account [152, 84, 106, 135], the label propagation algorithms can be regarded as random walk algorithms.

There have been attempts to integrate multi-view learning techniques into the label propagation algorithms. Academic communities developed the propagation learning algorithms around two ideas about graph construction:

1. Constructing a graph that is comprised of vertices and edges coming from different views. In literature, the choices for graph construction includes hyper-graph [38, 138], heterogeneous graph [59, 102, 31], and $\mathcal{K}$-partite [29] (e.g., bi-partite [51]) graph.
2. Fuse multiple similarity graphs into one graph. The fusion can be as simple as a linear combination of similarity graphs [127] or formulated as an optimization problem [69, 37, 9].

CHAPTER 3
CompoundEyes: Near-duplicate Detection in Large Scale Online Video Systems in the Cloud

3.1 Overview

In recent years we have witnessed the proliferation of video content on the Internet. This growth was fueled by rapid advances in multimedia technologies, and the popularity of online video hosting and sharing services (e.g., YouTube, Yahoo! video).

The expansion of video content is accompanied by ubiquitous duplication. Wu et al. [125] showed that among Internet 13,129 videos, around 27% are near-duplicate. Therefore, efficiently identifying near-duplicate videos (NDVs) on a large-scale is a fundamental research goal, which can benefit the performance of video sharing and hosting services from many aspects. For example, by identifying the NDV copies, bandwidth utilization and storage management in video content distribution systems can be further optimized; by comparing the tags associated with NDVs, the spamming videos can be detected as well; furthermore, the detection of NDVs allows pirated copies to be identified (e.g., YouTube Content ID).

Presently, a standard way in practice to detect NDVs is based on keywords, tags, or associated descriptions. However, these textual data are less reliable in detecting NDVs than visual content. It is very common that identical video clips have different sets of associated tags, while clips with an identical set of tags can be significantly different in visual content. Therefore, a content-based NDVD (Near-Duplicate Video Detection) system [125, 109, 52]

is more desirable than the one based on texts. These NDVD systems, however, tend to use high-dimensional feature representations and complicated algorithms to seek good detection accuracy, thereby sacrificing efficiency for accuracy. This approach is not practical for large-scale NDVD applications. As reported in YouTube statistics [136], 300 hours of video clips are uploaded every minute. If an NDVD system is not efficient enough, the detection speed cannot catch up to the video uploading speed. However, building a practical NDVD system is challenging, due to the following two reasons:

- **The Complexity of Data**: Compared with other forms of big data such as records or logs, videos are more information-abundant and complicated. Therefore, using features to profile a video is not as effective as it does in content-based duplicate document detection. In the cloud, there are numerous modifications of video content to produce NDVs, for example, variations in encoding format or parameters, photometric variations, or frame insertion or deletion. Every feature discovered hitherto has its own drawbacks because a particular part of the information about video content has been discarded by this feature.
- **Detection Speed Requirement**: In order to cope with the sheer volume and increasing speed, a fast video detection system is necessary. However, this requirement is contradictory to the practice of using high-dimensional and composite feature representations to embody videos [109, 104] because the construction of these representations is exhaustive [109]. Consequently, it is generally conducted offline [109, 104].

Intuitively, high-dimensional and composite visual feature representations are more informative, thus more discriminative in near-duplicate video detection or retrieval, in spite of the time-consuming constructions. This is why recent research focuses on using high-dimensional feature design and feature fusion

[109, 104] to detect NDVs. However, in this thesis, using the information entropy concept, we demonstrate that composite feature representations are not necessarily more informative than a collection of simple representations. In addition, the dimensional growth may further reduce the informativeness. Accordingly, we shift the focus away from the design of an advanced representation into the design of the whole detection system. We design and implement an efficient yet accurate NDVD system, called CompoundEyes [20]. Our idea was inspired by the compound eyes of insects, which are made up of numerous small optical systems. Although an individual small optical system is weak by itself, they together form a comprehensible eyesight, allowing for an incredibly wide viewing angle and the detection of fast movement.

The design of CompoundEyes seamlessly integrates the ideas of multiple instance learning and the principles of the systems approach. In CompoundEyes, every video is represented as a bag of feature vectors of different types. Each type of feature vectors is generated by an independent component that applies a feature extraction and representation algorithm. Although individual components are relatively weak in terms of accuracy, together as a system, they could achieve satisfactory accuracy improvement. Meanwhile, the system efficiency is ensured because the algorithms utilized by these components are simple, fast, and adapted for parallelism exploitation. We adopted the CC_WEB_VIDEO [128] dataset to evaluate the performance of CompoundEyes. Compared with a similar work [147], the accuracy has been improved from 80% to 89%, with only 1.45 seconds average temporal cost for videos less than 10 minutes in length. In addition, we evaluate CompoundEyes against two BoWs (Bag of Words) -based and two CNN (Convolutional Neural Network) -based near-duplicate video detection systems in terms of accuracy. The experiment results show that the detection accuracy of CompoundEyes

is on par with the 19-layer VGGNet [107], and higher than those of the other three systems.

The contributions of our system can be further explained by the following aspects:

- **A Shifting of Detection Paradigm**: We apply a new philosophy for the design of NDVD systems, which employs multi-feature information fusion with well-coordinated classifiers instead of multi-feature fusion with a simple classifier. Based on the definition of the informativeness of video representation, we prove that theoretically, a sophisticated representation combining multiple features does not provide more information than a collection of simple features. Thus the latter approach does not guarantee higher accuracy than the former one.
- **Efficiency Improvement**: We use low-dimensional representations to achieve efficiency and scalability. Though the accuracy of using individual features with reduced dimensionality is affected, we apply the Multiple Instance Learning approach for information fusion and make the final detection result more accurate than state-of-the-art approaches. Moreover, we exploit the parallelism in our system to further accelerate the detection speed.
- **Implementation**: Our implementation of CompoundEyes along with the simplicity of input representations and native support of parallelism exhibits satisfactory performance in terms of both accuracy and detection efficiency.

3.2 Preliminaries

There is no standard definition of Near-duplicate Video (NDV) in literature. In this thesis, we adopt the most stringent and least subjective [85] definition proposed by Wu et al. [125], in which NDVs are videos of similar

visual content but have undergone various modifications such as illumination changes or caption insertion. Therefore the NDV detection is based on visual content rather than semantics.

3.2.1 Two-stage NDVD/NDVR

Near-duplicate Video Detection (NDVD) and Near-duplicate Video Retrieval (NDVR) are different in their objectives, but the underlying techniques are communal. In detection, the goal is to identify all duplicate pairs from the video dataset, whereas in retrieval, the aim is to locate the videos that are near-duplicate to the query video and position them appropriately. The typical process of content-based NDVD/NDVR systems is comprised of two stages: (1) feature extraction and description, (2) neighborhood construction.

Feature extraction and description

A video feature is a summary of information in visual content, which should preferably be stable and sufficiently distinguishable. The range where a feature is extracted may span globally across the whole video, such as the color distribution, or be localized to a region, such as interest regions.

Extracting features from a video is conducted on a frame-by-frame basis. For instance, to calculate the color distribution of a video, the color distribution of each frame is calculated first, then the average of them is taken as the color distribution of the video.

Descriptors are constructed to represent the extracted features quantitatively. Among numerous descriptors, histograms are widely adopted, to represent both global features (e.g., color distribution), or local features (e.g., SIFT, and BoWs).

Neighborhood construction

Owing to speed efficiency concerns, in NDVD/NDVR systems, videos are generally described with global representations (e.g., signatures) instead of a

sequence of pattern symbols. When the first stage ends, videos are summarized as a point in a multi-dimensional feature space. Ideally, NDVs should be adjacent, whereas different videos should be distant in this feature space. With a distance measurement, we are able to identify the near-duplicate videos to a video by constructing its neighborhood in the feature space. In other words, this neighborhood is a decision boundary. Those videos reside within the boundary are regarded as duplicated videos to the given video, and others are non-duplicate.

As the size of dataset grows, the construction of neighborhood of videos is critical for detection speed improvement. As mentioned in Chapter 2, to accelerate this construction, storage, and retrieval assistance schemes such as hash tables, inverted indexing file, or LSH (Locality Sensitive Hashing) [61], are introduced. Converting video representations into hash codes expedites the similarity comparison between them; thus it is particularly helpful to the retrieval of K-nearest neighboring videos.

3.2.2 Feature-centered detection paradigm

Conventionally, the feature representation construction in the first stage is the core of NDVD system design, and it has been deeply studied by academic communities. In this part, we commence our discussion about this feature-centered detection paradigm with a theoretical model, upon which the drawbacks of this paradigm are investigated, to introduce and justify the design philosophy of CompoundEyes.

Mathematical Model

First, we define four relevant concepts in NDVD systems as follows.

Definition 1. The neighborhood of a video $v \in V$ is $U(v) = \{v' \in V | v' \in duplicate(v)\}$.

Definition 1 is independent of feature representations.

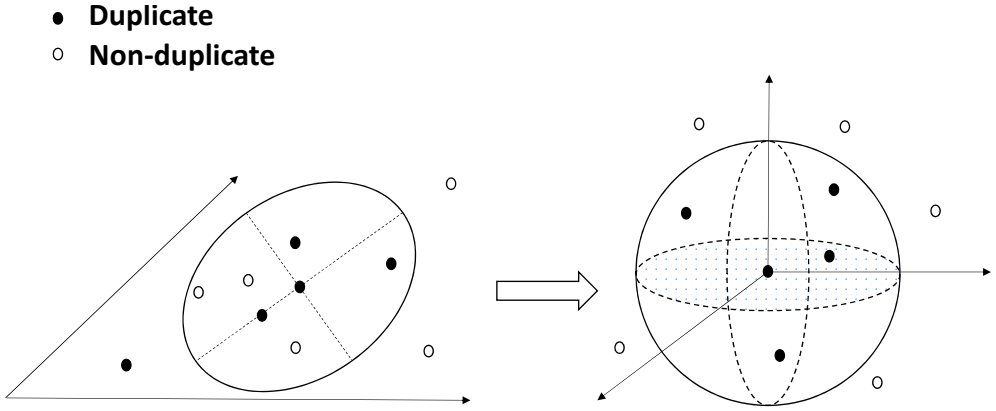


Figure 3–1: Transformation of Feature Space

Definition 2. The representation of a video $v \in V$ under feature $f \in F$ is defined as $X_f(v) \in R^n$ (i.e., Euclidean space). Defining the feature representation space as an Euclidean space is not mandatory.

Definition 3. The hypersphere neighborhood of a video $v \in V$ under feature $f \in F$ is defined as $S(X_f(v), \tau) = \{X_f(v') | v' \in V, |X_f(v') - X_f(v)| \leq \tau\}$, where $|.|$ is a distance measurement in the feature space.

Definition 4. The error set of $S(X_f(v), \tau)$ is defined as $E_f(v) = \{v' \in V | v' \in U(v),\ X_f(v') \notin S((X_f(v), \tau^*)\} \cup \{v' \in V | v' \notin U(v),\ X_f(v') \in S((X_f(v), \tau^*)\}$, where τ^* is the optimal value for $S(X_f(v), \tau)$.

With these definitions, after establishing the feature f, the classification task (e.g., NDVD/NDVR) in this paradigm is as simple as testing whether $v' \in S(x_f(v), \tau^*)$, $v, v' \in V$. Its accuracy can be measured by the volume of $E_f = \{E_f(v) | v \in V\}$. The smaller it is, the better f is to embody videos.

As shown in the left part of Figure 3–1, the hypersphere neighborhood in a simple, low-dimensional feature space under f_1 may not be a satisfactory approximation, as $|E_f| = 4$. To increase the discriminative ability, in the feature-centered paradigm, a higher-dimensional feature representation $X_f, f \in F$ is created by combining feature representations $X_{f_1}, X_{f_2}, \ldots, X_{f_n},\ f_1, f_2, \ldots, f_n \in$

F in various ways [104, 109, 110]. The hypersphere neighborhood in the feature space under f should be more accurate as shown in the right part of Figure 3–1, where $|E_f| = 0$. However, this paradigm may encounter issues from the following perspectives of dimensionality and informativeness.

Dimensionality

The first potential issue of the feature-centered paradigm is the high dimensionality of representations. Typically, there are two manners of dimensionality growth: more feature representations being integrated, or the vocabulary of visual words expanding. They can be illustrated with examples.

The LBP-based spatiotemporal feature [104] is an example of feature fusion. First, each frame is represented by a binary vector of 16 dimensions. Thus there are $2^{16} = 65536$ possible distinctive vectors, or patterns. Then the video representation, a histogram, is constructed by counting frames that fall into each pattern. In this way, the dimensionality of representations is 65536.

In BoWs methods, the dimensionality of representations is the number of visual words in the vocabulary, or $O(\sqrt{n})$ according to a rule of thumb, where n is the number of interest regions extracted from all videos. Given that there are 10^7 videos in a database, each of them has 10^2 frames and the average number of extracted regions is 10^3, the dimensionality of this representation, is $10^{\frac{7+2+3}{2}} = 10^6$.

Either the combinatorial explosion or sublinear growth could lead to the high-dimensionality of representations, which imposes heavy processing cost, thus reducing the detection speed of NDVD/NDVR systems. On the other hand, the accuracy could also be negatively affected. When dimensionality increases, the maximum distance between two random representations becomes

indiscernible compared to the minimum distance, as

$$\lim_{d\to\infty} E(\frac{dist_{max}(d) - dist_{min}(d)}{dist_{min}(d)}) = 0. \tag{3.1}$$

Thus the neighborhood becomes less meaningful. In addition, when more irrelevant or noisy dimensions are involved, the accuracy of neighboring video retrieval will also drop.

Informativeness

The second potential issue of the paradigm comes from the reduction of informativeness, which is critical to the detection accuracy. We assume that the feature representations emerge in the form histogram because it is widely adopted in describing both global features (e.g., color distribution) and local features (e.g. SIFT, PCA-SIFT, BoW). The informativeness of representation is defined as entropy.

Definition 5. Suppose $f_1, f_2, \ldots, f_k, \ldots \in F$ are visual features. Under each one of them, a video v can be represented as $X_{f_1}(v), X_{f_2}(v), \ldots, X_{f_k}(v), \ldots$. The informativeness of a video representation $X(v) \in \{X_{f_1}(v), X_{f_2}(v), \ldots, X_{f_k}(v), \ldots\}$ is $H_v(X) = -\sum_i p_v(x_i) \log p_v(x_i)$.

The term $p_v(x_i)$ in this equation is defined as $p_v(x_i) = g_v(x_i)w_i$, where $g_v : range(X) \to [0, 1]$ is the probability density function of $X(v)$, and w_i is the width of the i-th bin in this discrete probability density function. In other words, the range of $X(v)$ is divided into non-overlapping bins, and the value of $X(v)$ stays the same within a bin.

Formally, the definitions of these variables are: $w_i = u_i - l_i$, $x_i \in [l_i, u_i]$, $u_{i-1} = l_i$, $\cup_{i=1}^{n}[l_i, u_i] = range(X), i = 2 \ldots n$, where n is the dimensionality (i.e., number of bins) of $X(v)$.

The following properties regarding information lost could be revealed with Definition 5.

Property 1. *$H_v(X) = 0$, if $n = 1$; $H_v(X) \to 0$, if $n \to \infty$.*

Proof. By Definition 5, the proof of the first part is straightforward.

For the second part, as $n \to \infty$, $w_i \to 0$, thus $p_v(x_i) = g_v(x_i)w_i \to 0$. Additionally, according to the definition of entropy, $p(x)\log p(x) = 0$, when $p(x) = 0$. Therefore, $H_v(X) = -\sum_i p_v(x_i)\log p_v(x_i)_v \to 0$, as $n \to \infty$.

□

According to Property 1, increasing the dimensionality of representation does not necessarily make it more informative. On the contrary, as it becomes sparse, its informativeness is closer to 0. The attempt to construct more compact BoWs representation [64] buttresses this corollary. In essence, it reveals the curse of dimensionality as Equation 3.1 does, from another perspective.

Property 2. $H(X_{f_1}(v), X_{f_2}(v), \ldots, X_{f_k}(v)) \geq H(X_{f_i}(v)), i = 1, \ldots, k, \ldots$.

Property 3. $H(X_{f_1}(v), X_{f_2}(v), \ldots, X_{f_k}(v)) \leq H(X_{f_1}(v)) + H(X_{f_2}(v)) + \ldots + H(X_{f_k}(v))$;

Proof. $H(X_{f_1}(v), X_{f_2}(v), \ldots, X_{f_k}(v))$ is mathematically sound, because $X_{f_k}(v)$ can be viewed as a random variable. By Defintion 5, $P\{X_f(v) = x_i\} = g_v(x_i), f \in \{f_1, f_2, \ldots, f_k \ldots\}, x_i \in range(X_f(v))$. In this proof, we will regard $X_f(v)$ as a random variable instead of a feature representation in the form of histogram, and the definition of informativeness remains the same.

According to the chain rule,

$$H(X_{f_1}(v), X_{f_2}(v), \ldots, X_{f_k}(v)) = \sum_{i=1}^{k} H(X_{f_i}(v)|H(X_{f_1}(v), \ldots, H(X_{f_{i-1}}(v)). \tag{3.2}$$

Equation 3.2 can also be written as,

$$\sum_{i=1}^{k} H(X_{f_i}(v)|H(X_{f_1}(v),\ldots,H(X_{f_{i-1}}(v)) =$$

$$\sum_{i=1}^{k-1} H(X_{f_i}(v)|H(X_{f_1}(v),\ldots,H(X_{f_{i-1}}(v)) + H(X_{f_k}(v)). \tag{3.3}$$

Due to the non-negativity of entropy and Equation 3.3, the following inequality holds,

$$H(X_{f_1}(v), X_{f_2}(v),\ldots,X_{f_k}(v)) \geq H(X_{f_k}(v)). \tag{3.4}$$

Based on Inequality 3.4 and the symmetry of joint entropy, Property 2 can be proved.

To prove Property 3, we use the monotonicity property that conditioning reduces entropy,

$$H(X_{f_i}(v)|H(X_{f_1}(v),\ldots,H(X_{f_{i-1}}(v)) \leq H(X_{f_i}(v)), i = 1,\ldots,k. \tag{3.5}$$

By plugging Equation 3.5 into Equation 3.2, Property 3 can be proved. □

In Property 2 and 3, the joint distribution of random variables $X_{f_1}, X_{f_2},\ldots,X_{f_k}$ describes the essence of the approaches that combine these feature representations. From Property 2, constructing a sophisticated representation via feature fusion does increase its informativeness compared with every single feature representation. However, according to Property 3, the informativeness of this fused representation is upper bounded by the sum of the informativeness of the simpler representations. Therefore, building a sophisticated classifier, and feeding it with multiple feature representations could achieve higher accuracy than the combination of a fused representation and a simple classifier (i.e., the hypersphere neighborhood).

3.3 System Design

According to Property 1 and 3, we can realize the gains in accuracy by shifting the focus away from building an advanced feature representation towards an advanced classifier. In order to achieve fast detection speed, our system is designed according to the principles of the systems approach. Components are simple, efficient, and independent of each other. Parallelism generated from this autonomy is also exploited to increase speed further. In addition, efforts have been made to organize the feature extractors and classifiers to ensure satisfactory performance both in accuracy and speed.

3.3.1 Architecture

CompoundEyes is designed by using an abstraction layer model. In this model, frames are sampled at the Frame layer, in which features are extracted and represented at the Feature layer. From these representations, patterns of NDVs rest at the Knowledge layer, which finally emerge at the Decision layer and are used to make predictions about videos being duplicated or not.

The system is divided into three subsystems: Feature Vector Builder, Vector Repository, and Ensemble Learner. These subsystems are located on the Feature, Knowledge and Decision layers, as shown in Figure 3–2.

In all the related systems, most of the computational overhead is originated from the Feature Vector Builder subsystem. The subsystem is intrinsically complicated due to the complexity of the visual content of multimedia objects. By following the principles of systems approach, we divide the Feature Vector Builder subsystem into various Vector Builders, each of which uses a unique feature extraction and representation algorithm. For each Vector Builder, there is a weak Learner which uses its representations to make predictions. These predictions are collected by the Ensemble Learner, to make

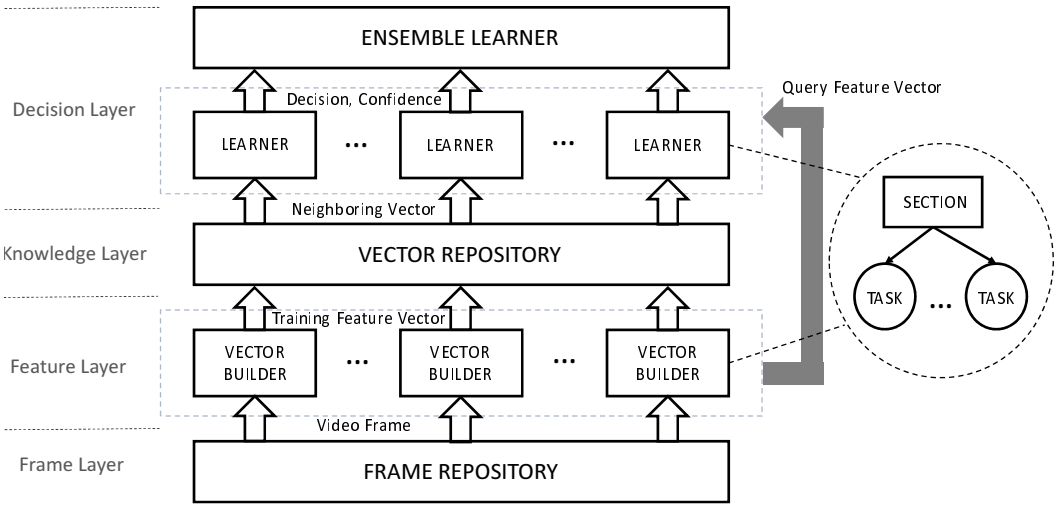


Figure 3–2: The Architecture and Parallel Organization of CompoundEyes

final predictions. This design also conforms to the ideas of multiple instance learning.

The division of the functionalities of the system ensures the exploitation of the hidden parallelism. The parallel organization of CompoundEyes is hierarchical, as illustrated inside the dashed rectangles and circle of Figure 3–2. The first level is the function parallelism among components, i.e., Vector Builders and weak Learners. They compete for parallel sections to perform their computations. The second level is the data parallelism within the computations of Vector Builders. Upon obtaining a parallel section, one or more parallel tasks are spawned, among which the computations of the Vector Builder are divided.

3.3.2 Data flow

Feature layer

In the Feature layer, we utilize seven broadly used feature extraction algorithms: color coherence, color distribution, LBP (Local Binary Pattern), edge orientation, ordinal pattern, motion orientation, and bounding boxes of objects, as explained in Figure 3–3. All of these algorithms are simple

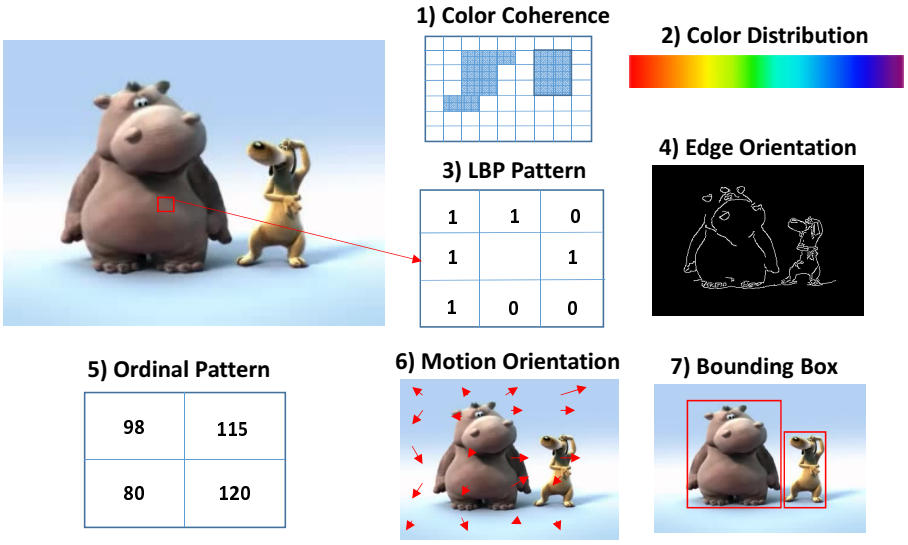


Figure 3–3: The Seven Features in CompoundEyes

and efficient. Furthermore, feature diversity is positively correlated with the accuracy of the final prediction [30].

All the Vector Builders work on a frame-by-frame basis. Given the j-th Vector Builder deals with feature $f_j, j = 1, \ldots, 7$, it first extracts f_j from the i-th key-frame of video v and represents it as a histogram $x_i^{f_j}(v), i = 1, \ldots, N(v)$, where $N(v)$ is the number of key-frames in v. Then the video representation of v built by this Vector Builder is calculated as $X_{f_j}(v) = \frac{1}{N(v)} \sum_{i=1}^{N(v)} x_i^{f_j}(v)$. The frame-level data parallelism in this calculation is exploited by distributing the computations of $x_i^{f_j}(v), i = 1, \ldots, N(v)$ onto the tasks belonging to a parallel section obtained by this Vector Builder, as shown in Figure 3–2. Although the sequential order of frames is ignored in computing the average of the feature representations of frames, its negative impact on the discriminative capability of the system can be offset by the employing of the motion feature.

Knowledge layer

To explain the neighborhood construction or neighboring video retrieval procedure in Vector Repository, we need the following definition.

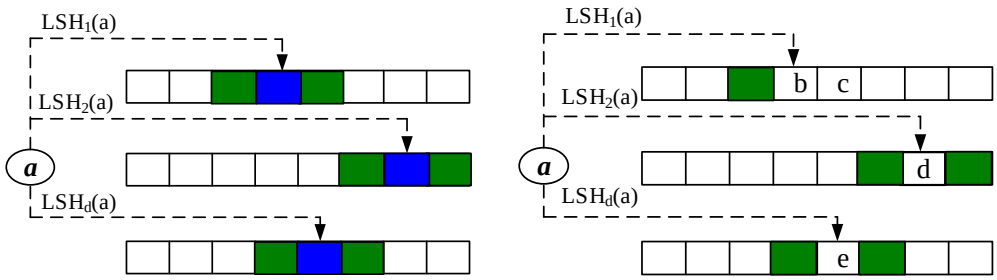
Definition 6. The neighborhood of a video $v \in V$ under feature $f \in F$: $U_f(v, \tau) = \{X_f(v') | v' \in V, |X_f(v') - X_f(v)| \leq \tau\}$.

After videos are represented as bags of vectorial representations $\{X_{f_j}(v) | j = 1, \ldots, 7\}, v \in V$, the representations of the videos in the training set V_t are stored and indexed in Vector Repository along with their ground-truth labels, separated by the features $f_j, j = 1, \ldots, 7$ into seven subspaces. This Vector Repository grants CompoundEyes the capability to act both as an NDVD or NDVR system.

When the representations of a query video v_q from the testing set V_q, $X_{f_j}(v_q), j = 1, \ldots, 7$ are issued to the Vector Repository, its neighborhoods under feature f_j, $U_{f_j}(v_q, \tau), j = 1, \ldots, 7$, are computed and returned to Learners in the Decision layer respectively. These videos are stored in the Vector Repository during the pre-processing and are attached with the ground-truth labels. The primary objective of the Video Repository is to make this neighboring video retrieval procedure more efficient.

The Vector Repository is organized as a NEST structure [57, 96], as shown in Figure 3–4(a), which is an LSH structure in essence. The advantages of our design are twofold. First, LSH is sensitive to locality thereby having the capability of providing the neighboring video retrieval with more accurate results. Second, the temporal cost of retrieval is $O(1)$. In addition, the LSH structure is combined with Cuckoo Hashing [97]. As a result, the problems of unbalanced load among hash tables and of local similar sets are mitigated, further enhancing its retrieval performance.

The blue and green buckets are potential neighbors to the vector in a Nearest Neighbor query. The blue ones are hit in an LSH computation, and their adjacent green ones also exhibit a correlation to the query. For example, in Figure 3–4(b), $LSH_1(a)$, $LSH_2(a)$ and $LSH_3(a)$ have been occupied by b, d



(a) A Multi-choice LSH Indexing Structure

(b) Available Locations for the Item a

Figure 3–4: The Indexing Structure of NEST[57]

and e, a is placed into an empty green bucket. Otherwise, if all the candidate positions are occupied, the "kicking out" operation in cuckoo hashing needs to be carried out to make room for the item.

The efficiency of the structure for a Nearest Neighbor query mainly depends on whether the inserted items are well-organized. When the load of hash tables is unbalanced, or items in the searching path form a local similar set, the performance of query would be deteriorated. To further enhance the performance of NEST, we modify the algorithm of insertion operations.

Denote $MaxLoop$ as the maximum kicking-out count, initialized to 0. At first, when the kicking-out count is under $MaxLoop/2$, we use the random selection to select an insertion position and record the count of position occurrences. Next, when the kicking-out count is between $MaxLoop/2$ and $MaxLoop$, the potential position with the minimum frequency is picked out for the next "kicking out". Afterward, random cuckoo hashing starts to make room for the item to be inserted.

The second step is to address the unbalanced load problem. For example, assume there are 3 hash tables, I, II and III, of size 10, and $MaxLoop$ is 5. The occurrence of a position becoming candidate along a kicking-out path is counted. Suppose item t will be inserted into the indexing structure, and its candidate positions are all occupied. The kick-out operation has experienced

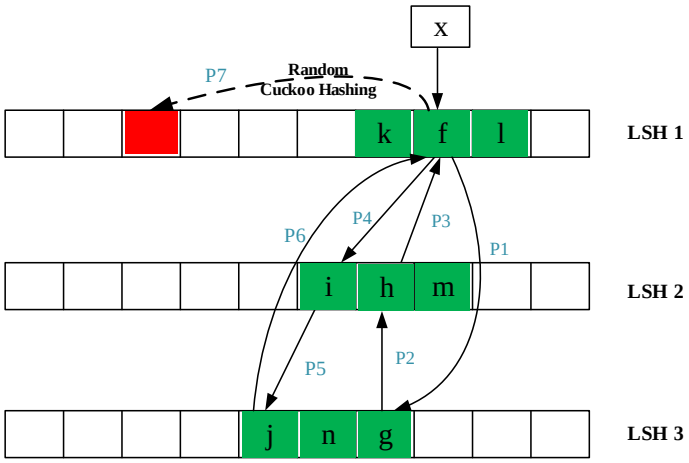


Figure 3–5: The Random Cuckoo Hashing in the Last Step.

the path $e \rightarrow c \rightarrow a$, whose candidate positions are $[(I,4)(II,5)(III,7)]$, $[(I,3),(II,5),(III,7)]$, and $[(I,3),(II,4),(III,7)]$ respectively, where $(II,7)$ is used to represent the 7th position in table II. When the kicking count reach to 3, greater than $MaxLoop/2$, the candidate positions of item a with the minimum occurrence, $(II,4)$ is chosen to be the next position to kick out.

The third step is introduced to jump out the local similar sets. As shown in Figure 3–5, assume item x will be inserted, and all the items marked are similar. Since the candidate and adjacent positions of x are all occupied, the position where item f stations are selected to start the kicking-out process. Let $P1, P2, ..., P6$ mark the kicking-out operations along the kicking path. We observe that two kicking circles are formed because LSH calculations form similar items into a similar set. In kicking circles, items kick each other, which eventually fails the insertion operation. Thus, we need to use the random cuckoo hashing to jump out the similar sets.

Decision layer

We model the NDVD task as a classification problem. A video $v \in V$ can belong to n possible classes $c_i, i = 1, \ldots, n$. For example, when $n = 2$, the classes are duplication and non-duplication. In CompoundEyes, $n = 7$,

because the dataset we adopt divides videos into seven categories: Exactly Duplicate, Similar, Different Version, Major Change, Long Version, Dissimilar, and Do not Exist. Dissimilar and Do not Exist are treated as the same.

As in Figure 3–2, the learners (or classifiers) in the Decision layer are organized in a hierarchical manner. The prediction of a video being duplicate is made upon the hypotheses of the seven weak Learners.

The weak Learners are denoted as $L_j, j = 1, \ldots, N$, where $N = 7$ is equal to the number of features we adopt. The videos from both the training set V_t and test set V_q are summarized as bags of representations $\{X_{f_j}(v)|v \in V_t \cup V_q, j = 1, \ldots, 7\}$ in the Feature layer. $\{X_{f_j}(v)|v \in V_t, j = 1, \ldots, 7\}$ are stored in the Vector Repository along with their ground-truth labels $\{v = c_i|v \in V_t, i = 1, \ldots, 7\}$, while $\{X_{f_j}(v)|v \in V_q, j = 1, \ldots, 7\}$ are directed to Learners $L_j, j = 1, \ldots, N$, respectively, as shown in Figure 3–2. On L_j, the probabilities $p(v_q = c_i|L_j), i = 1, \ldots, 7$ are approximated with frequencies,

$$Fn(v_q = c_i|L_j) = \frac{|\{v = c_i|v \in V_t, v \in U_{f_j}(v_q, \tau)\}|}{|\{v|v \in V_t, v \in U_{f_j}(v_q, \tau)\}|}, i = 1, \ldots, 7.$$

The computation of $U_{f_j}(v_q, \tau)$ is performed by the Vector Repository, as mentioned above.

These frequencies are taken as input to the Ensemble Learner, which calculates the posterior probabilities $p(v_q = c_i|L_1, \ldots, L_7), i = 1, \ldots, 7$, utilizing the BKS (Behavior-Knowledge Space) method [113] as follows,

$$p(v_q = c_i|L_1, \ldots, L_7) \cong \hat{p}(v_q = c_i|L_1, \ldots, L_7),$$

$$\hat{p}(v_q = c_i|L_1, \ldots, L_7) = \frac{Fn(v_q = c_i|L_1, \ldots, L_7)}{\sum_j Fn(v_q = c_j|L_1, \ldots, L_7)}.$$

To make estimating $Fn(v_q = c_i|L_1, \ldots, L_7), i = 1, \ldots, 7$ easier, , we assume $L_j, j = 1, \ldots, 7$ are conditionally independent, which is sensible because of

the diversity of features. With the approximation $p(v_q = c_i|L_j) \cong Fn(v_q = c_i|L_j), i = 1, \ldots, 7$, we have,

$$\begin{aligned} p(v_q &= c_i|L_1, \ldots, L_7) \propto p(L_1, \ldots, L_7|v_q = c_i) \\ &= \prod_{j=1}^{7} p(L_j|v_q = c_i) \propto \prod_{i=1}^{7} p(v_q = c_i|L_j) \\ &\cong \prod_{j=1}^{7} Fn(v_q = c_i|L_j), i = 1, \ldots, 7. \end{aligned}$$

Therefore, with appropriate normalization, the probabilities are estimated as

$$p(v_q = c_i|L_1, \ldots, L_7) = \frac{\prod_{j=1}^{7} Fn(v_q = c_i|L_j)}{\sum_{k=1}^{7} \prod_{j=1}^{7} Fn(v_q = c_k|L_j)}, i = 1, \ldots, 7.$$

The class with the largest posterior probability would be the final prediction of the class of v_q.

The combination of the Nearest Neighbor algorithm applied on the weak Learners, and the BKS method on the Ensemble Learner appears satisfactory to the design of CompoundEyes. First, the Vector Repository directly provides an interface to efficiently compute $U_{f_j}(v_q, \tau)$, whose cost is $O(1)$. Second, the Nearest Neighbor algorithm is non-parametric, which is helpful to reduce the training cost to $O(1)$, fulfilling the in-situ requirement. Third, the Nearest Neighbor algorithm is sensitive to the variations of feature types [30], thus making it suitable for the scenario of multiple feature subspaces. Fourth, the BKS method is sufficiently accurate to be applied to the Ensemble Learner [113].

3.3.3 Advantages

CompoundEyes can be used as both an NDVD and NDVR system because it not only determines whether a video is visually similar to another video but also calculate the similarity score (i.e., the posterior probabilities). The advantages of CompoundEyes can be illustrated by the following aspects:

Accuracy. The accuracy improvement is primarily achieved via the collective efforts of learners. First, the coverage of feature space is broader. Not only are spatial and temporal information used, but also color, edge orientation, texture, and object sizes information is also included in learning. Second, the diversity of representations enhances the accuracy of learning.

Detection Speed. Primarily, two factors contribute to the improvement of the detection speed. The first one is the compactness of representations, which shortens the temporal cost of extracting feature vectors in the pre-processing stage and of neighboring vectorial representation retrieval in the processing stage. The second one is the exploitation of the function parallelism among the Vector Builders and Learners, and the frame-level data parallelism within the Vector Builders.

In-situ Updating. CompoundEyes has the capacity of continually updating its classifiers when incorporates new knowledge (i.e., videos and corresponding ground-truth labels), because the cost of training classifiers is $O(1)$, and the changes in classifiers do not affect the construction of representations in the Feature Layer.

Modularity. The components in CompoundEyes are independent, and so can be changed without affecting others. For example, a new Vector Builder detecting a new type of features can be admitted if necessary, so is the case with weak learners implementing other algorithms, and the Vector Repository utilizing alternative indexing schemes. Therefore, the system could be easily upgraded.

3.4 Evaluation

3.4.1 Experimental setup

We implement CompoundEyes in C++, C, and Matlab. Specifically, Vector Builders are coded in C++, with the assistance of the OpenCV libraries.

Weak Learners and NEST are implemented in C, and the Ensemble Learner is programmed in Matlab. The parallel parts of CompoundEyes are implemented by using OpenMP libraries.

Experiments about CompoundEyes are conducted on a 64-core Intel Xeon E5-4640 machine (2.4GHz, 12.5GB memory) with Ubuntu system. The cores are distributed equally into 4 NUMA nodes. This multi-core machine is favorable for the parallel computing of CompoundEyes, which boosts the speed efficiency substantially.

CompoundEyes is evaluated against other NDVD/NDVR systems that adopt the CC_WEB_VIDEO dataset. The source code of these systems is not available, except MFH [109]. However, the demand for memory of the matrix computations in this system is too large to be satisfied by our machines. Therefore, in the comparisons of accuracy and response time, we adopt the values reported in the papers and compare them with those of CompoundEyes. With respect to the preprocessing time, the comparisons are conducted in a theoretical analysis manner. The reason is twofold. First, the time cost of the preprocessing of other systems is not provided in the papers. Second, the parallelization of the preprocessing of CompoundEyes would make the comparisons of experimental results unfair.

The Bag-of-Words (BoWs) feature representation is widely used by the vision community. Recently, the feature representations generated by deep neural networks have shown promising results in complicated tasks such as automatic image annotation. We implement two NDVD systems based on these two classic features and compare these systems with CompounEyes in terms of accuracy. The two systems are coded in Python, with the assistance of OpenCV and TensorFlow libraries. The BoWs-based system is deployed on a 4-core Intel i3-3220 machine (3.3GHz, 16GB memory), and the deep neural

network-based system is deployed on a 4-core Intel i5-4460 machine (3.2GHz, 12GB memory), whose GPU is GeForce GT-720. The reason for the change of machines is that we did not have access to the 64-core machine anymore when these experiments were conducted.

3.4.2 Dataset description

We evaluate CompoundEyes on the CC_WEB_VIDEO dataset. There are four reasons for this selection.

- First, it was constructed from real online videos. All the videos were downloaded from YouTube, Google Video, and Yahoo! Video.
- Second, various formats and editorial modifications are included.
- Third, it has been widely adopted, which facilitates us to compare the performance.
- Fourth, ground-truth labels are provided. These labels are obtained manually, which is laborious and makes the dataset precious for NDVD/NDVR research.

The CC_WEB_VIDEO dataset is comprised of 24 independent groups. In each group, a video is designated as the seed and others are compared with it and labeled accordingly. For instance, if a video has the same visual content as that of the seed video, it is labeled as "E"; if their contents are different, it is labeled as "X".

3.4.3 NDVD/NDVR systems in the literature

To evaluate the performance of CompoundEyes, we compare it with existing state-of-the-art NDVD/NDVR systems that have been evaluated on the CC_WEB_VIDEO dataset or on extended datasets. They are described as follows.

Hierarchical detection system (HIER): Wu et al. [125] proposed a hierarchical NDVD system, which uses a global signature-based method to

filter out duplicates with minor changes first, leaving more sophisticated ones to the local feature-based method.

Video Cuboid based detection system (VC): Zhou et al. [147] introduced the Video Cuboid signature, an n-gram based representation, to integrate the temporal and spatial information. Further optimizations include the use of the EMD distance, the incremental signature construction, and an LSH based matching scheme.

Spatial-temporal feature based detection system (ST): Shang et al. [104] explored alternative approaches to combining the temporal and spatial information into signatures. Two approaches are proposed: Conditional Entropy (ST-CE) and Local Binary Pattern (ST-LBP). The retrieval process is accelerated by applying a fast intersection kernel and inverted file.

Multiple feature hashing based detection system (MFH): Song et al. [109] provided another combination of a global and a local feature of videos. A series of hash functions are learned from feature representations. The neighboring video searching is conducted in Hamming space of the hash codes.

In these systems, VC provides us with the results of accuracy, while others are more concerned with mean average precision and average response time. Hence, we will compare CompoundEyes with VC in terms of accuracy, and with others in terms of mean average precision and average response time.

3.4.4 NDVD Systems based on classical visual features

These NDVD systems are designed in a quintessential feature-centered manner. The feature extraction and representation algorithms are advanced, through which the videos are represented as discriminative, high-dimensional vectors in the Euclidean space. The distances between the representations, along with corresponding ground-truth labels of videos in the training set are

fed into a one-vs-the-rest SVM (Support Vector Machine) classifier. When the training phase completes, this classifier is able to predict whether a video is duplicate or not based on vector distance. Because we compare CompoundEyes with these NDVD systems only in terms of accuracy, efficiency-boosting techniques such as LSH are not involved in the design.

Bag-of-Words (BoWs): As aforementioned, in the BoWs approaches, local features of an image are extracted first, then summarized into a global representation (i.e., a histogram of the frequencies of the occurrence of visual words). Converting a video into a BoWs representation consists of two phases: the construction of visual word vocabulary, and the interpretation of videos based on this vocabulary.

In the first phase, we extract two types of local features from frames and build vocabularies accordingly. The first type is SIFT (Scale-Invariant Feature Transform), and the second one is SURF (Speeded Up Robust Features). Both of them are effective for a variety of computer vision applications. We randomly select 10% of the local features of all the keyframes of the videos in the training set and perform K-Means clustering on them. By rule of thumb, the number of clustering centers is set to be the squared root of the count of selected local features.

Converting the videos in the test set into BoWs representations commences when the vocabulary has been built. Since the dimensionality of the BoWs representation equals to the size of the visual word vocabulary, the computation of the representation of a video can be conducted by merely adding the representations of each frame of the video.

Deep convolutional neural network (CNN): With large training sets, deep convolutional neural networks are capable of outperforming humans in visual recognition tasks. The trained networks with good generalizability

can be used as a base network in transfer learning. The feature representations generated by these networks are more effective in vision tasks than simpler descriptions such as color histograms or BoWs representations, even on a different image set.

By detaching the last softmax layer, a standard deep convolutional neural network converts a frame into a high-dimensional feature vector. The feature representation of a video can be computed by averaging the feature vectors of its keyframes. Generally speaking, the feature representations of the videos generated in this way are effective if the pre-trained neural network performs well in annotating the frames of the videos. From preliminary experiments, we discovered that in terms of the annotation performance, VGGNet [107] is superior to Inception-v3 [111] on the frames of the CC_WEB_VIDEO dataset. Therefore, we use the implementation of the 16-layer, and 19-layer VGGNet implemented in TensorFlow for the NDVD task. The dimensionality of the feature representations of both networks is 1000.

3.4.5 Experimental results

In this subsection, extensive experiments are conducted to evaluate the performance of CompoundEyes. Datasets of various sizes are constructed by randomly selecting videos from the CC_WEB_VIDEO dataset. Unless stated otherwise, in each one of them, 50% are used as the training set and the other 50% as the test set.

Accuracy

Evaluation metrics.

- **Accuracy**: It is computed as $AC = \frac{n}{N}$, the portion of correct predictions in total results.
- **Mean Average Precision**: The Mean Average Precision (MAP) is computed by averaging the Average Precision (AP) of each group g,

Table 3–1: The Comparisons of Performance with other NDVD/NDVR Systems in Literature

SYSTEM	VC	HIER	ST-CE	ST-LBP	MFH	**Ours**
AC(%)	80	N/A	N/A	N/A	N/A	**89.2**
MAP(%)	N/A	95.20	95.30	95.00	95.40	**99.75**
RT (ms)	N/A	9600	3.7	3.6	N/A	**0.2051**
PMU	N/A	$O(k)$	$O(n)$	$O(n)$	$O(k^3n^3)$	$\boldsymbol{O(k)}$
TC	N/A	$O(kn^2)$	$O(kn)$	$O(kn)$	$O(k^3n^3)$	$\boldsymbol{O(kn)}$

as $MAP = \frac{1}{24}\sum_{g=1}^{24} AP_g$, $AP_g = \frac{1}{n}\sum_{i=1}^{n} \frac{i}{r_i}$, where n is the number of correct predictions , r_i is the rank of i-th correct prediction.

Results. CompoundEyes shows an improvement in the detection accuracy. It achieves a higher Accuracy than the VC system, 89.28% vs. 80%, and outperforms other NDVD/NDVR systems in Mean Average Precision, as shown in Table 3–1.

In order to evaluate CompoundEyes against the two NDVD systems based on the BoWs and CNN feature representations, a subset of the CC_WEB_VIDEO dataset with 10% randomly selected video clips is constructed. The construction of a smaller dataset is because of performance considerations. The construction of the BoWs visual word vocabularies will fail due to the shortage of memory if more portions of videos or local features are involved in the computations of vocabulary construction. In addition, the time cost of the computations of the CNN feature representations for the videos is high, especially when these computations are conducted on the outdated machines.

Table 3–2 shows the comparisons of Accuracy and Mean Average Precision between CompoundEyes with the two classical feature-centered NDVD systems. Depending on what type of local features are extracted, or the number of layers in the convolutional neural network, the two systems can be

Table 3–2: Accuracy Comparison with Classical Feature-based NDVD Systems (Implemented)

SYSTEM	BoWs-SIFT	BoWs-SURF	CNN-16	CNN-19	**Ours**
AC (%)	79.27	78.66	76.93	78.80	**80.91**
MAP (%)	98.26	98.18	92.53	97.66	**99.21**

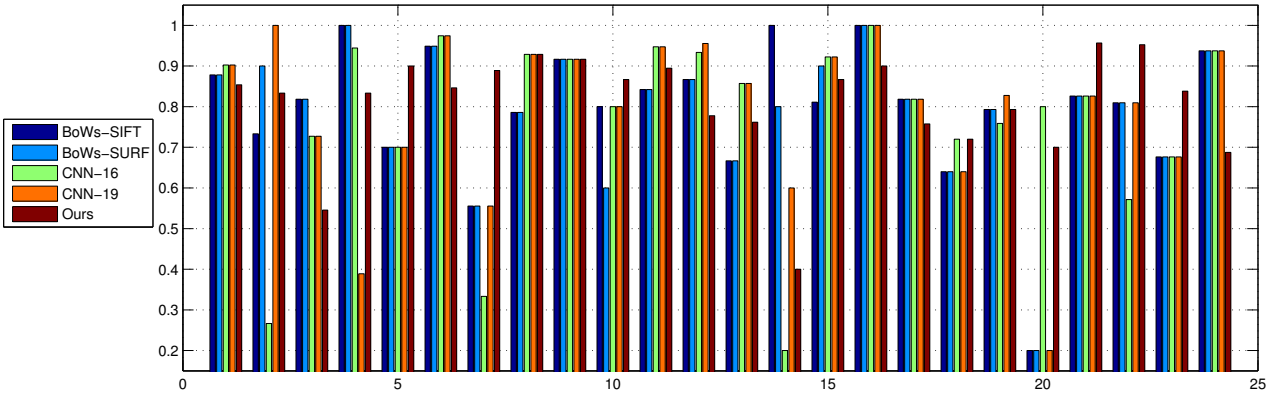


Figure 3–6: The Comparison of Accuracy with Classical Feature-based NDVD Systems on the 24 Groups of Videos

further divided into four systems (i.e., BoWs-SIFT, BoWs-SURF, CNN-16, CNN-19).

From Table 3–2, we can see that CompoundEyes is more effective than other NDVD systems based on BoWs and CNN features in terms of both Accuracy and Mean Average Precision, despite the simplicity and low dimensionality of the features that it applies, and the low cost of training. To further investigate the reason behind these counter-intuitive comparison results, we decompose the comparisons of average Accuracy and Mean Average Precision for all the videos in the subset into the comparisons of Accuracy and Average Precision over the 24 groups of videos. The results of the comparisons are shown in Figure 3–6 and Figure 3–7

Based on Figure 3–6, we count how many times that each of the five NDVD systems achieves the highest and lowest Accuracy in the comparisons

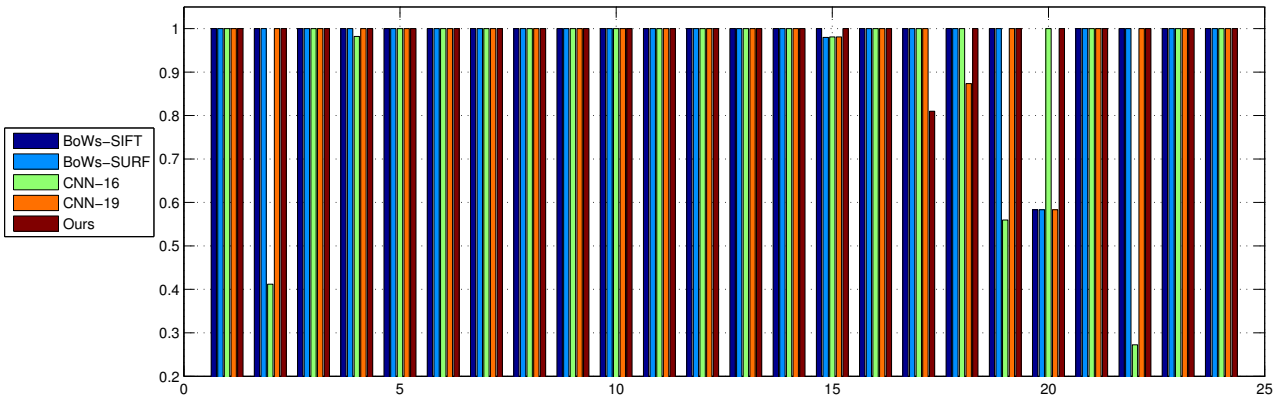


Figure 3–7: The Comparison of Average Precision with Classical Feature-based NDVD Systems on the 24 Groups of Videos

for each group. The results are summarized in Table 3–3. From this table, we notice that compared with the 19-layer VGGNet, CompoundEyes has fewer championships in the contest of the highest Accuracy. However, the Accuracy of the 19-layer VGGNet is less stable than CompoundEyes for the 24 video groups. For group 20, the Accuracy of VGGNet is as low as 20%, but CompoundEyes achieves 70%. Therefore CompoundEyes outperforms VGGNet in terms of average Accuracy.

The comparisons in Figure 3–7 are slightly different than the ones in Figure 3–6. Both the 19-layer VGGNet and CompoundEyes reach the 100% Average Precision for almost the 24 video groups, whereas other three systems fail for certain groups. In conclusion, from the comparisons of Accuracy and Mean Average Precision, CompoundEyes built on simple visual features surpasses or is on par with the sophisticated 19-layer VGGNet. The main reason that the deep neural network approach does not outperform CompoundEyes is that these networks are not fine-tuned, or trained to fit the CC_WEB_VIDEO dataset. However, it is not feasible to fine-tune a deep neural network for every task and every dataset.

Table 3–3: The Number of Times that the Five NDVD Systems Achieve the Best and Worst Performance in Terms of Accuracy

SYSTEM	BoWs-SIFT	BoWs-SURF	CNN-16	CNN-19	**Ours**
Highest	7	6	12	13	**9**
Lowest	11	11	9	7	**7**

Detection speed

The definition of temporal cost. The detection speed of CompoundEyes is measured by the temporal cost, which is the sum of the preprocessing time and response time.

$$Temporal\,Cost = Pre-processing\,Time + Response\,Time.$$

Analysis of preprocessing time. In literature, preprocessing is performed offline thus its temporal cost is not measured. The burden of preprocessing can be estimated from the fact that feature extraction of HIER, ST-CE or ST-LBP on a dataset of 132647 videos is practically impossible [109].

Suppose the number of videos is n, and the average number of keyframes in a video is k. The peak memory usage and worst case time complexity of the pre-processing of various systems are estimated in Table 3–1.

According to Table 3–1, CompoundEyes has advantages in both the peak memory usage and time complexity. It neither involves the computations and pairwise comparisons of SIFT descriptors as HIER, nor the computations of global variables, for example, the entropy of ordinal relations in ST-CE, the correlation between LBP patterns in ST-LBP, and the transformation and bias matrices in MFH, which are both spatially and temporally exhaustive. In contrast, the two major operations of CompoundEyes in preprocessing, the construction of representations, and inserting them into the hash tables which

can also be regarded as the training process, are both spatially and temporally efficient. The average temporal cost of preprocessing in experiments is 1.4537s, 99% of which is the cost of building feature vectors.

Experimental results of response time. The advantage of CompoundEyes in detection speed can also be manifested from response time, as shown in Table 3–1. The average response time of CompoundEyes only accounts for 5.70% of ST-LBP's.

Implementing the main part of CompoundEyes in C++, rather than Matlab may contribute to the reduction of response time. However, such a substantial reduction could not be explained merely by the efficiency of C++. In CompoundEyes, the dimensionality of representations could be 16, 32, or 64, all of which are much lower 65536 of ST-CE and ST-LBP [104]. This reduction in dimensionality is the main reason for the improvement on response time.

Parallel speedup

Experiments in this subsection are also performed on a 10% subset of CC_WEB_VIDEO. To evaluate the parallel speedup, the temporal costs of sequential version and parallel version are compared.

The average temporal cost of each Vector Builder for computing the feature vector of a video is estimated in Figure 3–8 first, and used as a reference for workload distribution. On the horizontal axis are the abbreviations of the features they extract, which are color histogram (HSV), color coherence (CC), ordinal pattern (SP), edge orientation (EO), bounding boxes of objects (BB), local binary pattern (LBP), and motion orientation (OPT_FLOW).

Thread Allocation Strategies. Both of the parallel sections and tasks in Figure 3–2 are abstractions of threads. Under different thread allocation strategies, the overall parallel speedup would be different. Therefore, we design

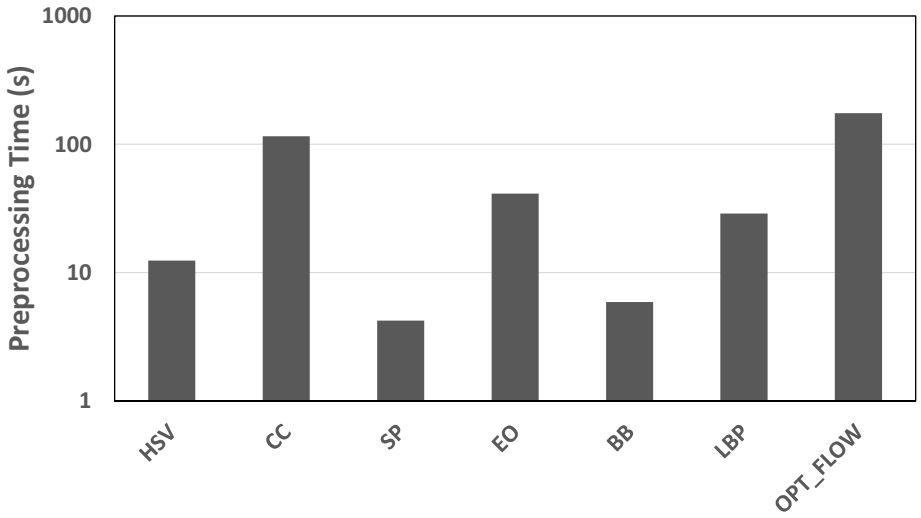


Figure 3–8: The Sequential Pre-processing Time of All the Vector Builders.

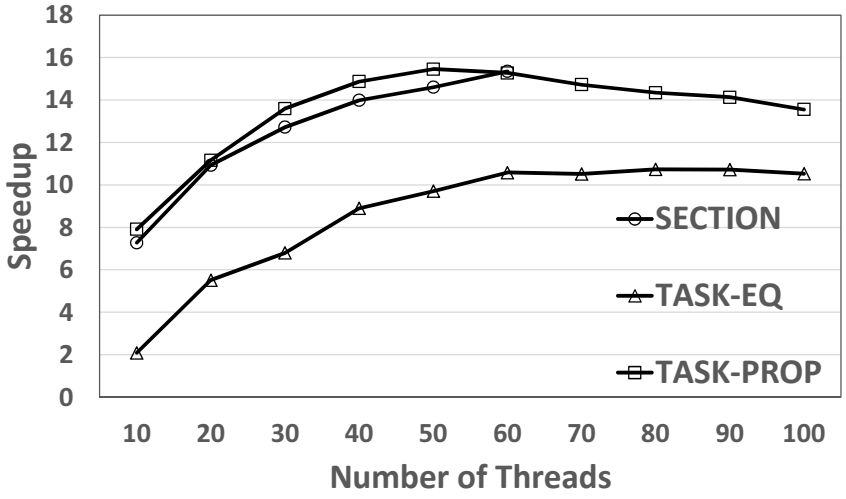


Figure 3–9: The Speedup of CompoundEyes Under 3 Thread Allocation Strategies

and compare three allocation strategies as follows, to sensibly provision the computing resources:

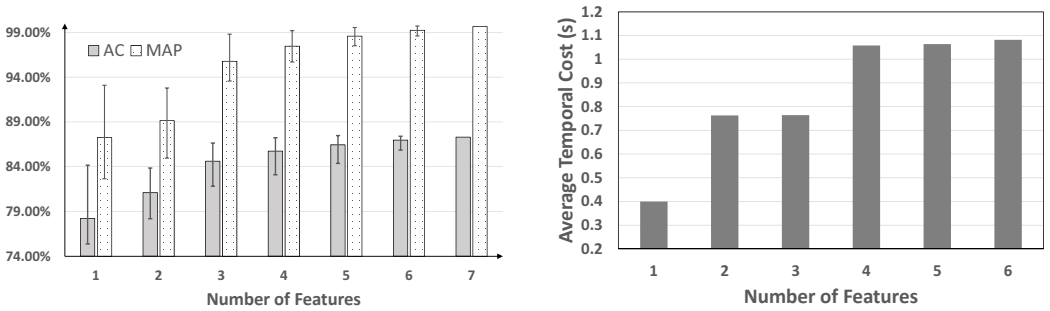
- **SECTION**: What varies in this strategy is the number of parallel sections competed by Vector Builders, from 1 to 7. Once a parallel section obtained, a number of parallel tasks will be allocated for computing. This number is proportional to the Vector Builder's sequential running time.
- **TASK-EQ**: In this strategy, every Vector Builder acquires a parallel section. What varies is the number of tasks spawned by a section, which is same for all the Vector Builders.
- **TASK-PROP**: In this strategy, not only does every Vector Builder obtain a parallel section, but also the number of tasks allocated to a Vector Builder is proportional to its sequential running time.

Results. As expected, from Figure 3–9, TASK-PROP achieves the best speedup, because it efficiently utilizes allocated threads. Moreover, we notice that when the thread number exceeds 60, the increase of speedup ceases. This value coincides with the number of cores in the machine. This phenomenon is a hint of resource contention.

We also notice that even under the best thread allocation strategy, the speedup is far from linear speedup. This is determined by the fact that in CompoundEyes, videos are processed sequentially, which limits the throughput of the system.

Feature information fusion

In this part, we assess the impact of feature information fusion, mainly on the detection accuracy. The experiments are conducted on a 10% subset. For the sake of fairness, the number of parallel sections is equal to the number



(a) On AC and MAP (b) On Temporal Cost

Figure 3–10: The Effect of Feature Information Fusion

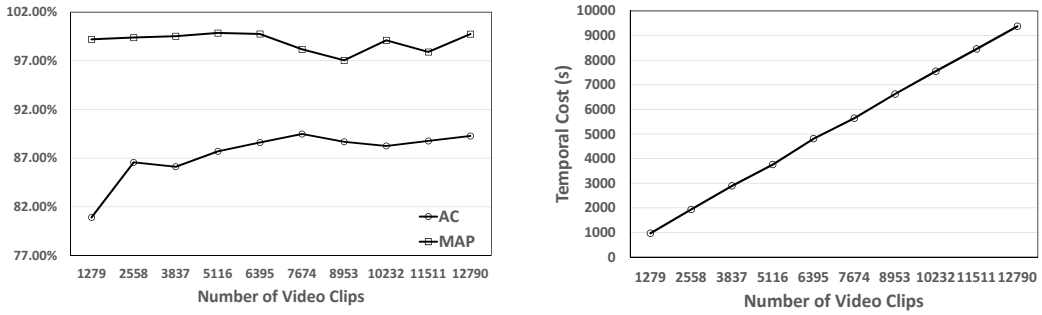
of features to be used, and the number of tasks that a section can spawn is equal for all the Vector Builders.

As shown in Figure 3–10(a), on average, the fusion increases the detection accuracy, both in terms of Accuracy and Mean Average Precision. This advantage becomes smaller when measured by the best accuracy of fusion. For example, the accuracy difference between the optimal combination of three features and four is negligible. This suggests the importance of the selection of feature information to be fused.

For the optimal combinations except all-included, corresponding average temporal costs are shown in Figure 3–10(b). They are helpful when choosing the number of features. For example, fusing three is better than four, because it costs less time but achieves comparable detection accuracy.

Relevant parameters

The Scale of the Dataset. The first relevant parameter is the scale of the dataset. According to Figure 3–11(a), the Accuracy is above 80% when the size is 1279, which is satisfactory. It also increases as the size of dataset grows. Therefore, CompoundEyes is accurate when sufficient knowledge has been learned, and its discriminative capability develops as knowledge accumulates. On the other hand, the MAP fluctuates after the size grows beyond 6395. This



(a) On AC and MAP (b) On Temporal Cost

Figure 3–11: The Effect of the Dataset Scale

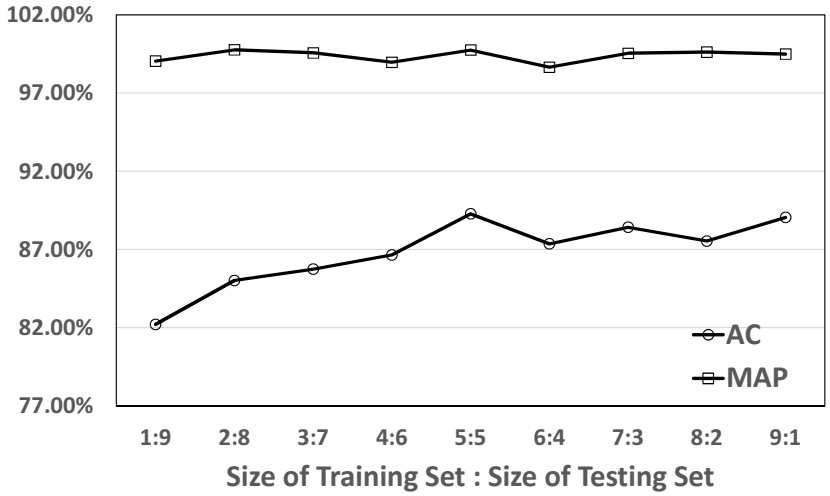
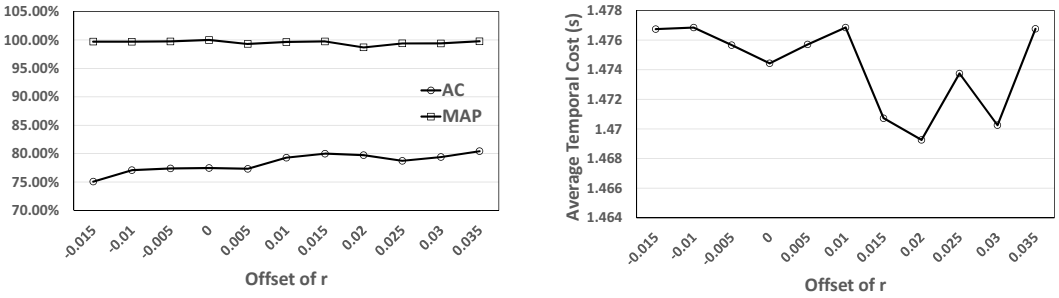


Figure 3–12: The Effect of the Portion of the Training Set on AC and MAP

phenomenon shows that it becomes more difficult to rank the videos as the size of the database increases.

Figure 3–11(b) affirms that the total temporal cost increases linearly rather than exponentially with the growth of dataset. This linearity confirms that Vector Repository is capable of maintaining decent performance even if the size of dataset becomes large.

The Portion of the Training Set. Because a system well-tuned on the training set could behave poorly on the test set, it is necessary to evaluate the detection accuracy of CompoundEyes under different portions of the training set.



(a) On AC and MAP

(b) On Temporal Cost

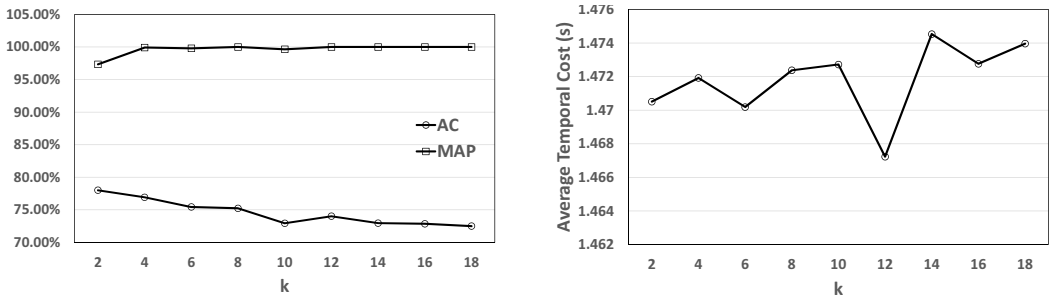
Figure 3–13: The Effect of r

The effect of this portion on Accuracy and Mean Average Precision is recorded in Figure 3–12. The value of MAP stays stable, and the value of AC increases as the ratio increases. Both of them peak around $5:5$. Afterward, more training videos do not enhance the detection accuracy the classifiers.

NEST-related Parameters. Two NEST-related parameters, r and k, are of importance. Parameter r is used as τ in the definition of the neighborhood in feature space. Parameter k is the number of hash tables. Generally speaking, a larger value of k increases the detection accuracy, at the expense of longer response time.

Because the value of r is different for each type of feature representations, we set them by experience first, then change them with the same offset. The effect of r on Accuracy and Mean Average Precision is shown in Figure 3–13(a), and the effect on average temporal cost is shown in Figure 3–13(b).

Since k is same for all feature spaces, we vary its value directly. From Figure 3–14(a), we observe that Accuracy and Mean Average Precision exhibit different trends, the former one goes down, and the latter one goes up and stays around 100%. This is because as k increases, the recall of neighboring feature representation retrieval grows, but the precision goes down. These changes reflect on Accuracy but not Mean Average Precision, for the number of correct results, and their ranks are barely affected.



(a) On AC and MAP (b) On Temporal Cost

Figure 3–14: The Effect of k

The effect of k on average temporal cost is shown in Figure 3–14(b), from which we know that 12 is the optimal value for the detection speed.

CHAPTER 4
An Empirical Study of the Textual Content of Web Videos

4.1 Overview

In the era of big data, both scientific and industrial communities benefit from the mining and analysis of the data that they process, for example, Internet search logs, meteorology, genomics, or financial records. In contrast to these numerical or textual data, multimedia data has been staying relatively low-key in the high-profile and incentive data business, which is not proportional to its volume. This phenomenon does not imply that multimedia data is petite or has been overlooked. In fact, there have been attempts to change this situation. For example, Prism Skylabs, a video analytics company, helps retailers to predict the preference of customers from camera recordings [49]. Instead, it is the complexity of the data itself that keeps it away from the center of the stage. Generally, online multimedia data consists of at least two significant aspects, visual and textual content, each of which requires high-dimensional vectors to describe. Meanwhile, these contents are not independent, their relationships make them more challenging to analyze.

Of the two contents, textual content plays a more important role than visual content does in industrial communities because textual words are a primary form of human knowledge, and it is relatively more efficient to process text than visual content. Therefore, despite the progress that has been made in recent years, the searching of photos or videos in real-world search engines are reliant on textual content, e.g. titles, tags, time-sync tags, descriptions, or comments. Although the ranking algorithms of search engines such

as YouTube remain covert to the public, plenty of users have inferred the significance of textual content and summarized tips to manipulate the rankings by modifying it [1].

Academic communities went further in exploring the relationships between visual and textual content. Their goals are automatic recognition and annotation of multimedia data. By feeding large neural networks with millions of photos, computers can be trained to recognize the visual content of photos [2]. In order to accomplish such a great task, photos in the training set should be accompanied by verified descriptive pieces of text. In real-world applications, this textual content is acquired in a decentralized rather than centralized style. Users across the world collectively contribute to the interpretations of visual content from their own perspectives. The absence of a quality control authority raises the following issues:

- **Sparsity**: The length of the accompanying text of a multimedia item (e.g., photo, video) is generally short. It is difficult to comprehensively present visual content in such a short piece of text.
- **Inaccuracy**: Textual content can be inaccurate, because of the intrinsic complexity of natural languages (e.g., synonymity, polysemy), the disparity in views and habits, or malicious spamming behaviors [3].

Besides the quality issues of textual content, the semantic gap between the low-level visual features (e.g., color, shape, texture), and the high-level human understanding expressed in text poses a huge challenge to every researcher in this field. A variety of methods, for instance, optimization [115], semantic latent models [133], or generative probabilistic models [33, 98], have been developed to narrow down the gap and improve the performance of applications such as video clustering [115], automatic video annotation [98, 133], and cross-modal tag cleansing [33]. In these methods, textual content is generally

summarized into latent topics or categories, which is then connected with the topics or categories extracted from the visual content. The inaccuracy in textual content is treated as noises that make the topic model stray. However, due to the absence of an empirical, preliminary study on the textual content of online multimedia data, the validity of these methods is in question. In this paper, we conduct such a study on the CC_WEB_VIDEO dataset by using CompoundEyes developed in Chapter 3. Our contributions in this Chapter could be summarized as follows.

- **Properties of Textual Content**: We studied the composition of the textual content of online videos, and the statistical distribution of the English words in it. We observe that even under the influences of numerous factors (e.g., user preferences), the frequency distribution of words follows Zipf's Law distribution, which is a signature of complex systems and social nature. In addition, we confirm the properties of sparsity quantitatively.
- **Hypotheses Validation**: The CC_WEB_VIDEO dataset is public dataset for the research of NDVD/NDVR (Near-Duplicate Video Detection/Retrieval). Conventionally, these tasks are performed on the visual content, but in this thesis, we conducted similar experiments on textual content as well. Through these two parallel experiments, the hypotheses about the relationships (or coherence) between them can be revealed. Moreover, we test whether the occurrence of URLs implies video spamming.

4.2 Similarity Measures of the Visual and Textual Contents

4.2.1 Visual similarity measure

Visual content-based NDVD is relatively slow and prone to inaccuracy. To increase the detection speed, while preserving decent detection accuracy, we designed an open architecture of NDVD/NDVR system, CompoundEyes,

as shown in Figure 3–2. Various types of visual features are extracted from the shared frame repository and represented as relatively low-dimensional vectors by the Vector Builders. These vectors are further inserted into the hash tables in the Vector Repository, where each table collects a type of feature vectors. Weak learners in the decision layer learn patterns from the corresponding hash tables and pass this knowledge on to the Ensemble Learner, to make the final decision about the similarity of a video v_i to v_j, which can be denoted as $sim_v(v_i, v_j)$.

4.2.2 Textual similarity measures

As aforementioned, the length of the accompanying text of online videos is short. Thus the methods for paragraphs similarity estimation is not suitable. In this thesis, we use three typical sentence similarity measures [4] to estimate the textual similarity.

Jaccard similarity coefficient

The Jaccard similarity coefficient belongs to word overlap similarity measures. The similarity score is computed based on the number of words shared by the two sentences:

$$sim_{jaccard}(s_1, s_2) = \frac{|\{s_1\} \cap \{s_2\}|}{|\{s_1\} \cup \{s_2\}|},$$

where $\{s_i\}, i = 1, 2$ is the set of words in the sentence s_i.

TF-IDF vector similarity

The TF-IDF (Term Frequency-Inverse Document Frequency) similarity is a member of vector space based methods, which is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. The similarity score is defined as the cosine similarity between the vector representations of the sentences. For the sake of reducing the degree of the sparseness of representation, words appear in the sentences rather than the

whole text collection are indexed and used as the feature set. The calculation is as follows:

$$sim_{tfidf}(s_1, s_2) = \frac{vector(s_1) \cdot vector(s_2)}{||vector(s_1)|| \cdot ||vector(s_2)||}.$$

Semantic similarity

The sentence semantic similarity measure belongs to the family of linguistic measures. It combines semantic and syntactic information and is calculated as follows:

$$sim_{sem+wo}(s_1, s_2) = \alpha sim_{sem}(s_1, s_2) + (1 - \alpha) sim_{wo}(s_1, s_2),$$

$$sim_{sem}(s_1, s_2) = \frac{1}{2}\left(\frac{\sum_{w \in \{s_1\}} \left(argmaxSim(w, w') \times idf(w)\right)}{\sum_{w \in \{s_1\}} idf(w)} + \frac{\sum_{w \in \{s_2\}} \left(argmaxSim(w, w') \times idf(w)\right)}{\sum_{w \in \{s_2\}} idf(w)}\right),$$

$$sim_{wo}(s_1, s_2) = 1 - \frac{||r_1 - r_2||}{||r_1 + r_2||}.$$

$Sim(w, w')$ is the Lin similarity, $idf(w)$ is the TF-IDF value of w. Lin similarity calculated on the information content of the least common subsumer and that of the two input synsets. The term $sim_{wo}(s_1, s_2)$ is the word order similarity of sentences, taking word composition into account, where $r_i, i = 1, 2$ is the word order vector of sentence s_i. The details about representing a sentence into a word order vector can be found in [4]. In this thesis, this semantic similarity measure is used as a proxy to gauge the semantic similarities between sentences or paragraphs.

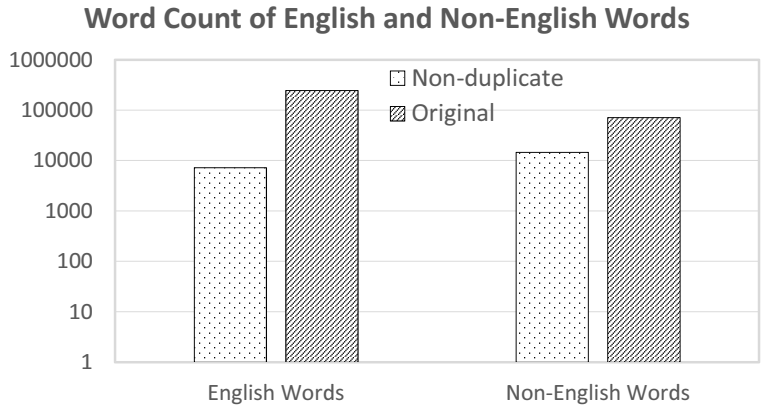


Figure 4–1: The Amount of English and Non-English Words

4.3 Statistical Properties of Textual Content

4.3.1 Composition

In the textual content of the dataset, there are special characters, URLs, and text. The special characters have been removed in the pre-processing because they are too deformed to decipher.

We divide the words into two categories: English and non-English, and count the quantities of words in each category, before and after the exclusion of duplication. The results are shown in Figure 4–1, from which we notice that on average, an English word has a higher probability of being used than that of a non-English word. This phenomenon suggests that English is the universal language of this dataset. Hence, we mainly focus on English words in this study.

4.3.2 Word frequency distribution: the Zipf's law

We plot the frequency (or occurrences) of words against the count of the words of that frequency in the log-log scale. As other large linguistic corpora, the frequencies of the words in CC_WEB_VIDEO exhibits a quintessential Zipf's Law distribution, regardless of whether they are from the "Title" (Figure 4–2(a)), "Tags" (Figure 4–2(b)), or "Description" (Figure 4–2(c)) field. It is surprised to see that the distribution of words in titles follows Zipf's Law

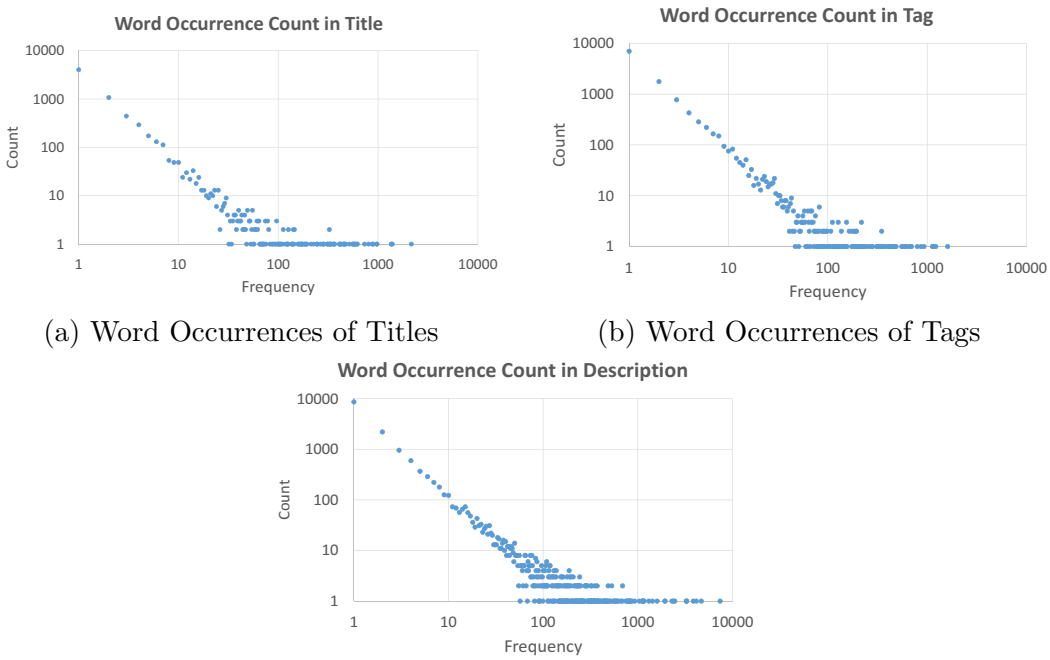
because the search keywords are specific thus the choices for titles are limited. The phenomenon suggests that the textual content of CC_WEB_VIDEO is sufficiently large and complicated to be treated as a complex system. The Zipf's Law distribution in such a system indicates the existence of hierarchy in words or the social nature of the textual content.

By comparing Figure 4–2(a), 4–2(b), and 4–2(c), we find that they follow a similar decaying trend, except that the tail portion of the distribution becomes denser as the average text length increases. Considering that the tail corresponds to frequently used words, this fact indicates that a longer piece of text such as description has a higher probability of containing these words.

The words used in the dataset can be divided into function words (i.e. words that have little lexical meaning or have ambiguous meaning, but instead, serve to express grammatical relationships with other words within a sentence, or specify the attitude or mood of the speaker, such as "the" or "he"), and content words (i.e., nouns, verbs, adjectives, and most adverbs). As shown in Figure 4–3(a), and 4–3(b), it is the content words, whose distribution is akin to the one in Figure 4–2(c). In contrast, the distribution of function words resembles the tail of Zipf's Law distribution, because the major portion of them are high-frequency words.

4.3.3 Sparsity

In literature, sparsity is a constantly mentioned property of multimedia annotation. Thus we want to observe whether it appears in the textual content of CC_WEB_VIDEO. The distributions of the length of titles, tags, and descriptions, in terms of word count, are drawn in Fig. 4–4, as a cumulative distribution function figure.



(a) Word Occurrences of Titles

(b) Word Occurrences of Tags

(c) Word Occurrences of Descriptions

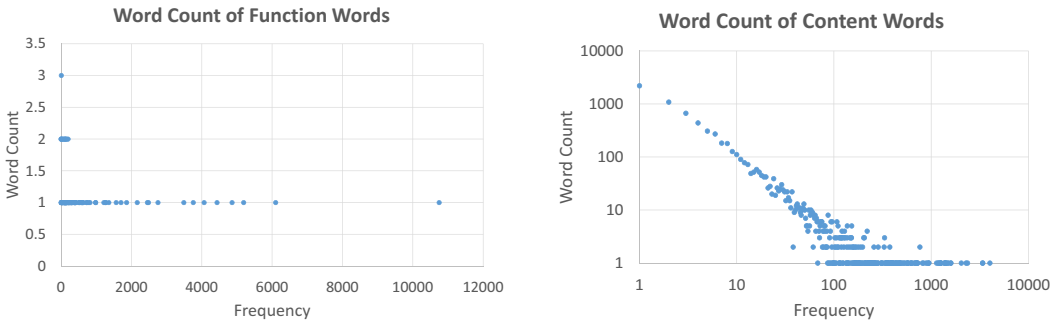
Figure 4–2: The Distribution of Word Occurrences of Textual Data. The X-axis Records the Frequencies of Appearance of Words in the Dataset, While the Y-axis Indicates How Many Words Have the Same Frequency of Appearance

From this figure, we find that the length of the textual content is indeed short. For titles and tags, at least 80% is less than 10 words. As for descriptions, 90% of them have less than 40 words. In such a short length, it is not feasible to elaborate the meaning of the visual content. Consequently, the vocabulary becomes more personal, which can negatively influence the correlation between visual and textual content.

4.3.4 Aggregating videos uploaded by the same user

According to the tripartite graph of folksonomy and its social nature, the preferences and habits of users may have an influence on textual content. Therefore it is necessary to aggregate the content on users and analyze the results.

In Figure 4–5, the distribution of the quantity of video clips uploaded by each user is depicted, which manifests the social aspect of online video sharing



(a) Function Words (b) Content Words

Figure 4–3: The Distribution of Word Occurrences of Function and Content Words. The X-axis Records the Frequencies of Appearance of Words in the Dataset, While the Y-axis Indicates How Many Words Have the Same Frequency of Appearance

and hosting websites. Like other social networks, we observe the existence of "super" users, who upload far more video clips than others, yet constitutes only a small portion of users. It is reasonable to highlight the influences of the "super" users in a model of textual content, but the problem is, are these "super" users trustworthy?

We traced the uploading records of the users whose names appear at least twice. The visual content, titles, tags, and descriptions of their uploaded videos are compared with those of the corresponding seed video. The comparison results of visual content are predominantly dissimilar, and those of the textual content show interesting features, which are shown in Figure 4–6.

In Figure 4–6, as the number of uploaded videos by one user increases, the similarity scores of titles and tags decrease, whereas the scores of descriptions grow to a high level (above 60%). In other words, as the number of videos uploaded by a user grows, the titles and tags of these videos become diverse, but the descriptions of them tend to be alike.

This phenomenon is strange, so we analyzed the textual content provided by the user who uploaded most videos, "rrichy92". Its characteristics fit exactly with what we describe above. The titles and tags are disparate,

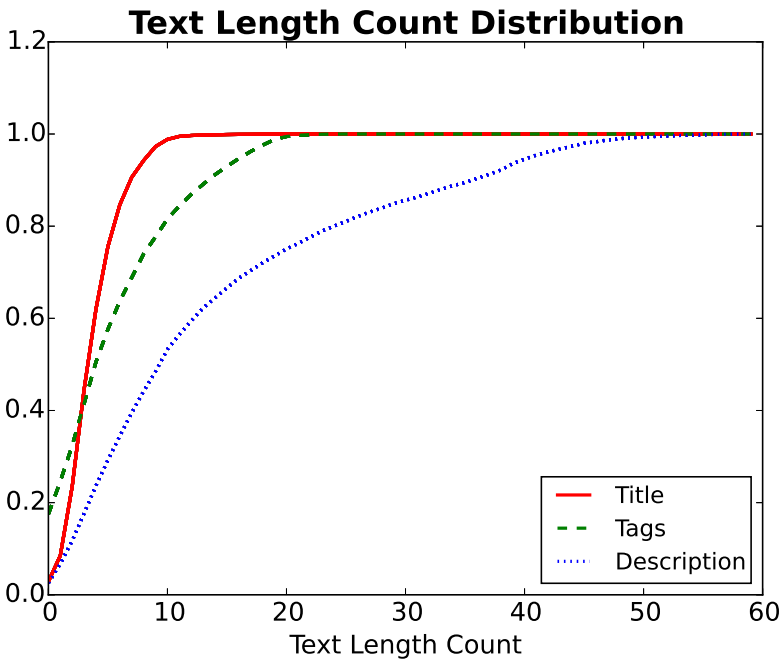


Figure 4–4: The Text Length Distribution of Title, Tags, and Descriptions

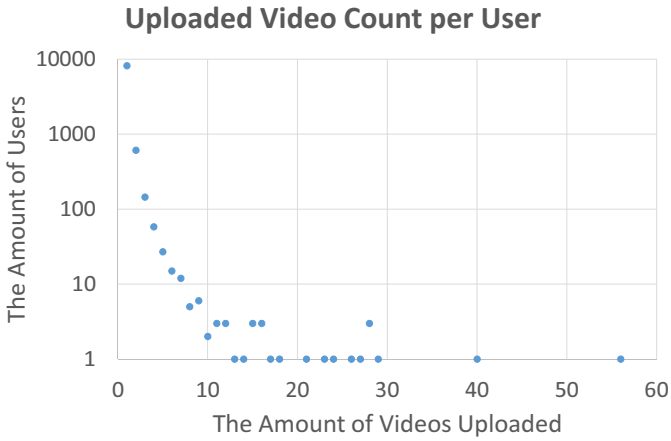


Figure 4–5: The Distribution of the Amount of Video Uploaded per User

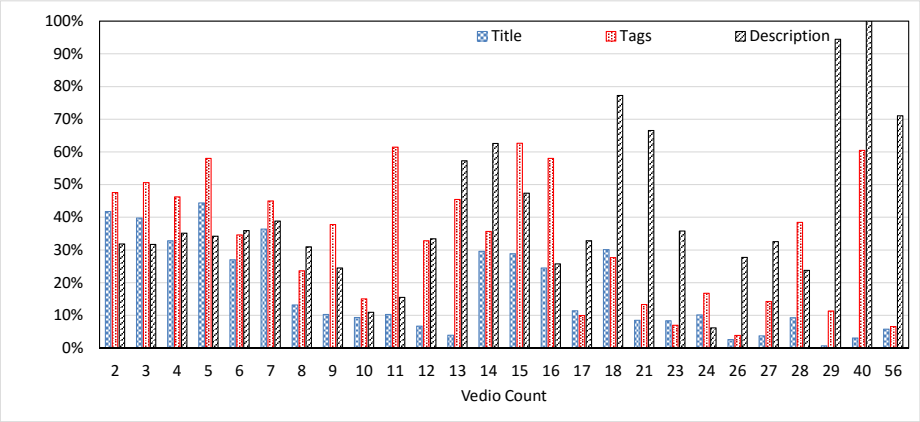


Figure 4–6: The Distribution of the Similarity Scores of the Videos Uploaded by the Same User over the Amount of the Videos

but every video description contains a sentence "MORE TAGS: Sex girl hot teens porn webcam xxx funny crazy yes genesis beatles daily show john jon stewart stephen steven colbert report tom green cruise". Evidently, this user was attempting to draw more attention from search engines, by adding the most searched keywords to all the descriptions. This extreme case shows the existence of video spammers.

4.4 The Quality of Information Retrieval

The quality of textual content has a considerable impact on the retrieval performance of search engines, thus measuring the quality of video retrieval results gives us an understanding of the quality of the textual part of the dataset.

The duplicate rate, or the percentage of near-duplicate videos in the dataset, on one hand, indicates the ubiquity of redundancy of visual content on the Internet; but on the other hand, is a measure of the relevance of search results to the query keywords. According to Wu et al. [125], around 27% videos in the dataset are duplicate or near-duplicate, in other words, 27% search results are relevant. More detailed statistics are shown in Figure 4–7,

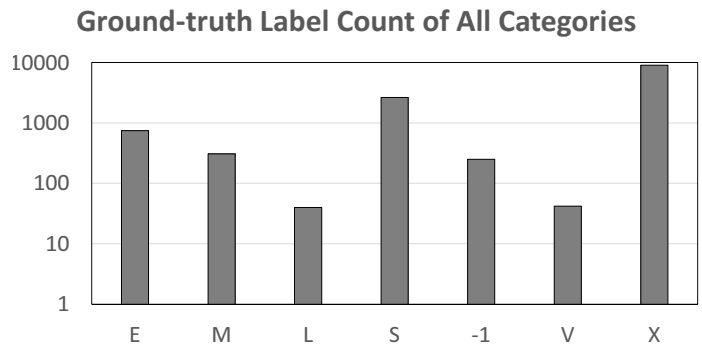


Figure 4–7: The Distribution of Ground-truth Labels over 7 Categories

from which we see that most of the videos are labeled with "X" or "-1" (i.e., "Dissimilar").

Various factors can lead to the less satisfactory performance of information retrieval. From the perspective of textual content, for example, the ambiguity of search keywords can widen the scope of the search. On account of this, we explored the impact of the generality of search keywords, by measuring the N-gram value with Google Ngram Viewer [45], the results are depicted in Figure 4–8. Contrary to expectation, from this figure, there is no significant relevance between the duplicate rate of videos and the N-gram value (or the generality of keywords).

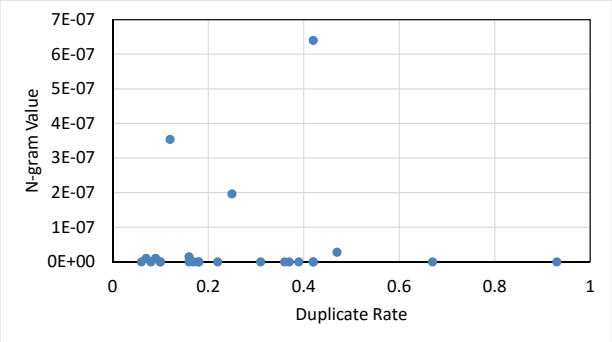
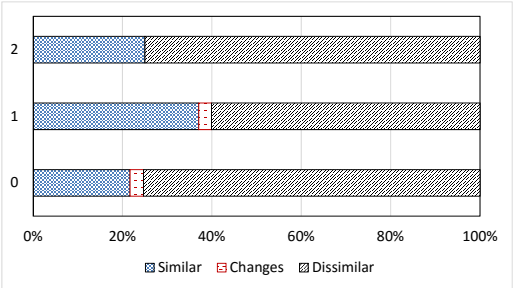


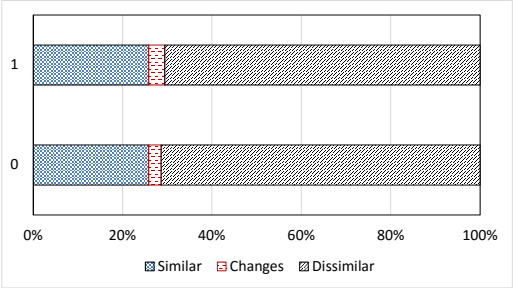
Figure 4–8: The N-gram Values of Keywords

Another case has also been studied. With more appearance of search keywords in titles, tags, or descriptions, corresponding videos should be more

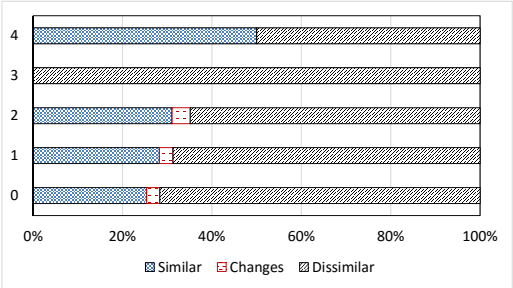
relevant and ranked higher in the search results. From Figure 4–9(a), 4–9(b), and 4–9(c), we notice that as the number of matches increases, the relevance rate of videos does increase as well, but not significantly. In addition, it rarely exceeds 50%.



(a) In Title



(b) In Tags



(c) In Description

Figure 4–9: The Matches of Keywords over the Categories of Search Results

4.5 URL: indicator of Video Spam?

The appearance of URL has been used as a heuristic feature of video spam [10] because one of the intentions of positioning URLs in the textual content field such as "Description" is to attract more traffic to the sites to which they point. In this dataset, the distribution of the frequency of URLs in the fields

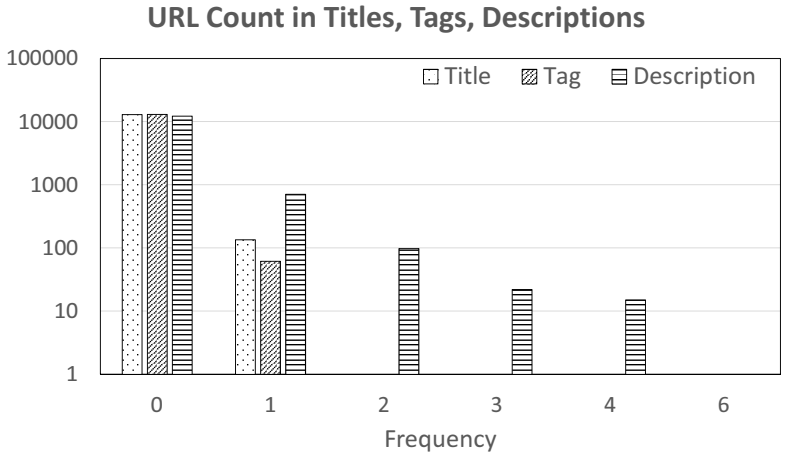


Figure 4–10: The Distribution of URL Occurrences in All Fields

Table 4–1: The Correlation between URL's Appearance and Visual Content Relevance

	Similar	Changes	Dissimilar
In Title	0.154412	0.036765	0.808824
In Tags	0.222222	0	0.777778
In Description	0.199528	0.042503	0.757969

"Title", "Tags", and "Description" of video clips is shown in Figure 4–10. From this figure, we discover that URLs rarely appear in the textual content of this dataset.

We also use Jaccard Coefficient to assess the frequency of URL's appearances in two textual content fields. For "Title" and "Tags", the coefficient is 0.058127, for "Title" and "Description", the value is 0.087432, and for "Tags" and "Description", it is 0.02941. All of them are quite small, which implies the rarity of the occurrences of the same URL in two fields of a video.

The scarcity of URL's occurrences and co-occurrences in textual content fields is not desirable for high recall rate to a heuristic feature of spamming. At least the spammer identified by aggregating videos on users mentioned above does not put URL in his/her description of videos.

On the other hand, the occurrence of URL may not be an accurate spammer indicator either. From Table 4–1, we see that the emergence of URL, whether in titles, tags, or descriptions, does increase the irrelevance rate of videos compared with average rate (73%), but the disparity is not significant.

4.6 Correlation between Visual and Textual Content

In this section, the similarity scores computed on visual content by CompoundEyes, constitute the visual content view. This view is compared to the textual content view, or the similarity scores calculated on titles, tags, and descriptions respectively. The experiments in this section use 50% of the dataset as the training set, the other half as the test set.

All the comparisons in the section are used to validate two following hypotheses, which are the two aspects of the homophily assumption:

- **Hypothesis 1:** similar visual content is accompanied with similar textual content;
- **Hypothesis 2:** similar textual content is accompanied with similar visual content.

4.6.1 Hypothesis 1

To validate this hypothesis, we divide the instances in the test set into three categories (i.e., "Similar", "Changes", and "Dissimilar") according to their NDVD detection results. The similarity scores of the textual content of these videos are computed field by field, under the three different textual similarity measures. The results are summarized in Table 4–2, 4–3, and 4–4.

In order to make the comparisons of absolute similarity scores meaningful, we conduct Wilcoxon signed-rank test along the rows and columns of these tables, on each pair of items. In Table 4–2, along the columns, when the Jaccard Coefficient or semantic similarity measure is applied, the values of category "Changes" and category "Dissimilar" are not significantly different;

Table 4–2: The Textual Similarity Scores on the Title Field

	Jaccard	TF-IDF	Semantic
Similar	0.571	0.668	0.449
Changes	0.308	0.357	0.331
Dissimilar	0.284	0.304	0.350

Table 4–3: The Textual Similarity Scores on the Tags Field

	Jaccard	TF-IDF	Semantic
Similar	0.332	0.414	0.360
Changes	0.217	0.272	0.280
Dissimilar	0.211	0.239	0.287

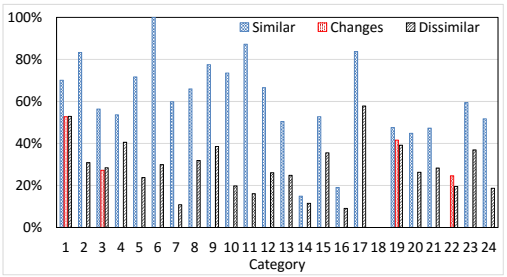
along the rows, for category "Dissimilar", the values of the Jaccard Coefficient and TF-IDF measures do not have a statistical significant difference. In Table 4–3, along the columns, the values of all the three measures for category "Changes" and "Dissimilar" do not have a statistically significant difference; along the rows, for category "Similar", the scores of the Jaccard Coefficient and semantic similarity measures are not significantly different; for category "Changes", the scores of all the three measures are not significantly different. In Table 4–4, along the columns, for category "Changes" and "Dissimilar", the values of all the three measures are not significantly different; along the rows, for category "Changes", the scores of the Jaccard Coefficient and TF-IDF measures do not have a statistically significant difference.

The similarity scores overall are lower than we expected, so we break down them according to which group (from 1 to 24) the videos belong. The results

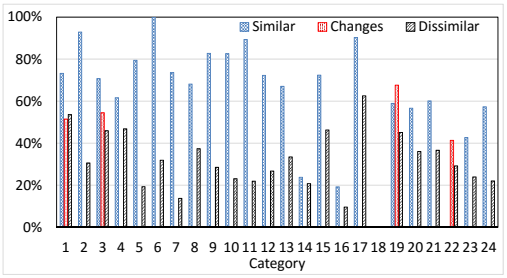
Table 4–4: The Textual Similarity Scores on the Description Field

	Jaccard	TF-IDF	Semantic
Similar	0.228	0.306	0.351
Changes	0.128	0.188	0.266
Dissimilar	0.120	0.163	0.299

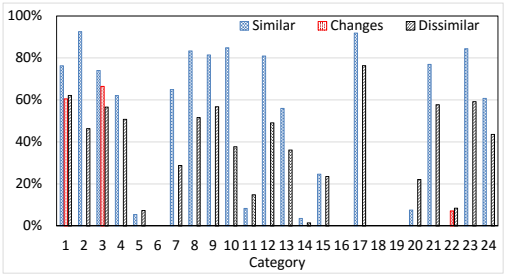
are shown in the following nine figures: Figure 4–11(a), 4–11(b), and 4–11(c) are computed using the Jaccard Coefficient score; Figure 4–12(a), 4–12(b), and 4–12(c) are computed using the TF-IDF similarity score; Figure 4–13(a), 4–13(b), and 4–13(c) are calculated using the semantic similarity score. The three figures in each set depict the similarity scores on "Title", "Tags", and "Description" fields respectively.



(a) Jaccard Coefficient

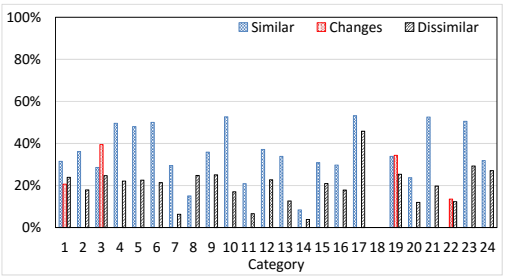


(b) TF-IDF

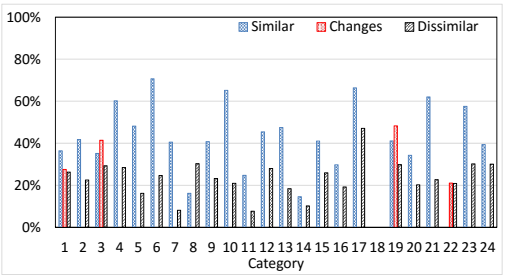


(c) Semantic

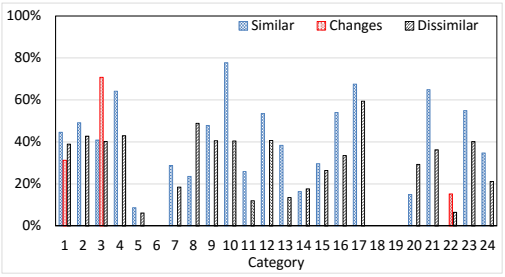
Figure 4–11: Similarity Scores of the "Title" Field of Videos of 24 Groups



(a) Jaccard Coefficient



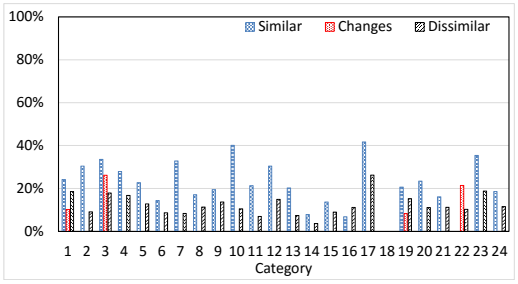
(b) TF-IDF



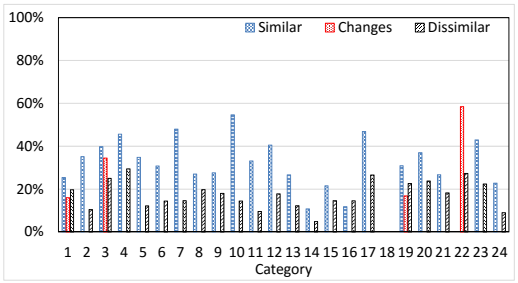
(c) Semantic

Figure 4–12: Similarity Scores of the "Tags" Field of Videos of 24 Groups

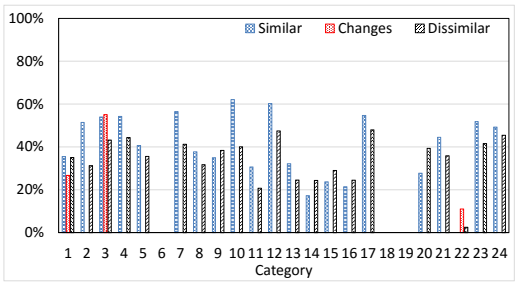
Generally speaking, the similarity scores of the videos belong to the "Similar" category are higher than those of the videos in the "Dissimilar" category.



(a) Jaccard Coefficient



(b) TF-IDF



(c) Semantic

Figure 4–13: Similarity Scores of the "Description" Field of Videos of 24 Groups

Thus Hypothesis 1 holds. However, there are outliers in these figures, for example, videos from group 22 in Figure 4–11(a), from group 20 in Figure 4–12(c) and 4–13(c), from group 16 in Figure 4–13(a), 4–13(b), and 4–13(c), from group 14 in Figure 4–13(c), from group 15 in Figure 4–13(c). The textual content of the videos in these groups is more stray and noisy. Meanwhile, the adoption of the semantic similarity score on the "Description" field has the largest probability of violating Hypothesis 1.

In all 9 figures, videos from group 18 are outliers. All the similarities scores are 0. This is because there are a lot of Chinese characters in the textual content of the videos. They are removed during the pre-processing.

4.6.2 Hypothesis 2

To validate Hypothesis 2, we divide the range of textual similarity scores, under different measures and for different fields, into 10 bins. For the videos in each bin, we depict the proportions of them in the three categories. The results are also presented in nine figures, which are organized in the same way as the above nine figures.

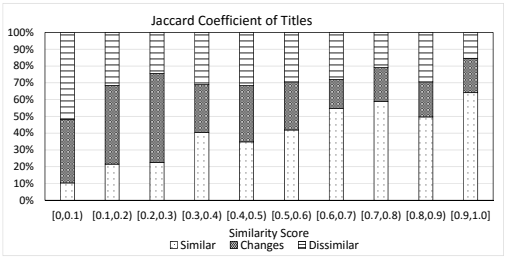
As Figure 4–14(a), 4–15(a), 4–16(a), 4–14(b), 4–15(b), and 4–16(b) show, the increase of textual similarity scores is positively correlated with the increase of the proportions of videos of "Similar" and "Changes". In these cases, Hypothesis 2 is valid. However, this trend becomes less distinguishable in Figure 4–14(c), 4–15(c), and 4–16(c).

4.6.3 Relevant factors

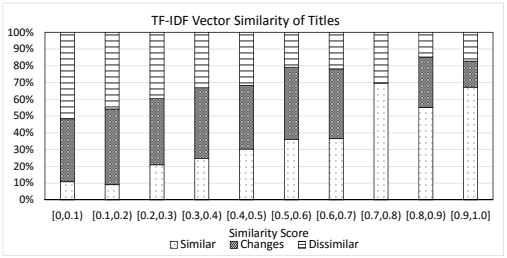
The effect of textual similarity measures

Along the rows of Table 4–2, 4–3, and 4–4 respectively, we can see that:

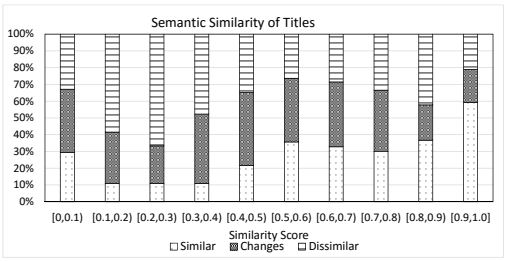
- TF-IDF vector similarity scores are the highest for the videos in the category "Similar".
- The Jaccard Coefficient scores are the lowest for those in the category "Dissimilar".
- For those in the category "Changes", the highest can be the TF-IDF vector similarity or the semantic similarity, the lowest are also the Jaccard Coefficient scores.
- When the ratio between the average scores of "Similar" and "Dissimilar" videos is considered (higher ratio implies higher discriminative ability of



(a) Jaccard Coefficient

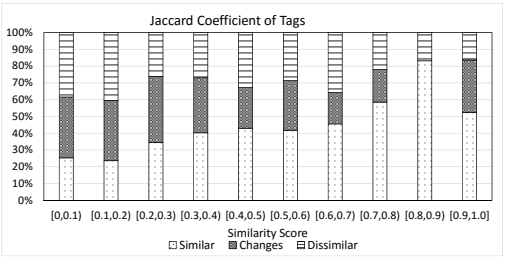


(b) TF-IDF

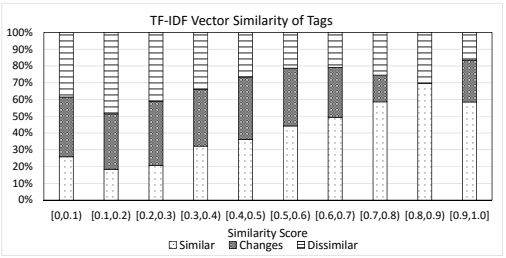


(c) Semantic

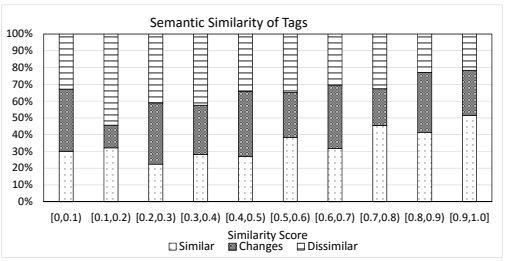
Figure 4–14: The Distribution of Visual Content Categories over the Similarity Scores of Titles



(a) Jaccard Coefficient



(b) TF-IDF



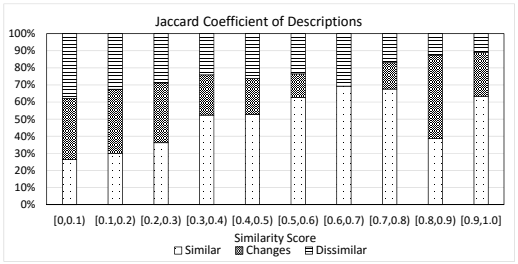
(c) Semantic

Figure 4–15: The Distribution of Visual Content Categories over the Similarity Scores of Tags

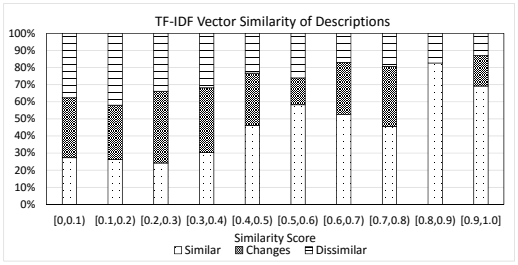
similarity measure), the ranking is the TF-IDF vector similarity, Jaccard coefficient, and semantic meaning, in descendant order.

Based on these observations, TF-IDF vector similarity is the best textual similarity measure out of the three. It correctly assigns high values for the videos in "Similar", low values to those in "Dissimilar", keeps a more discernible difference between them.

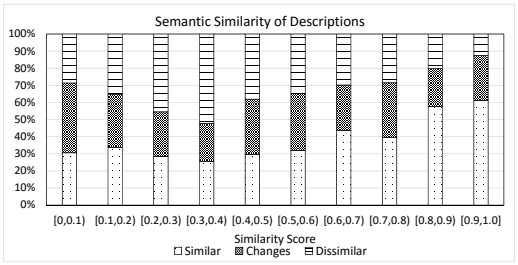
These textual similarity measures are evaluated and compared with others in paper [4]. Their performance depends on the complexity of dataset. In a



(a) Jaccard Coefficient



(b) TF-IDF



(c) Semantic

Figure 4–16: The Distribution of Visual Content Categories over the Similarity Scores of Descriptions

low-complexity dataset, linguist measures such as semantic similarity score are superior, but in a high-complexity one, TF-IDF and Jaccard Coefficient similarity scores are better. This conclusion is confirmed by the observations in this part. In addition, it proves the high complexity of the textual content of the dataset.

The effect of textual similarity measures

By rearranging Table 4–2, 4–3, and 4–4, highlighting the effect of the fields, we observe that:

- For almost every video category and every textual similarity measure, the score of the “Title” field is the highest and the lowest is that of the “Description” field.
- There is only one exception, which is the semantic similarity score for the descriptions of videos in category “Dissimilar”.

From these points, the ”Title” field is the most suitable field for the similarity computations on textual content, and the “Description” field is the least ideal field. The scores computed in this field are low and less distinctive.

CHAPTER 5
Cross-Modal Relevance Learning of Online Multimedia Data

5.1 Overview

Multimedia data has been regarded as the "biggest big data" nowadays [50]. Despite its widely acknowledged importance, multimedia data content remains less manageable, searchable, or reusable compared with other types of data. A potential solution to understanding the complex multimedia data is to annotate the data with the textual description, and the visual content of them can be referenced and managed on the level of semantic meanings.

Recently, deep neural networks, such as CNN (Convolutional Neural Network) or RNN (Recurrent Neural Network), have made startling progress in multimedia data understanding, and have been successfully applied in tasks such as image retrieval [124], automatic annotation of images [95, 71, 75, 90, 70]. The recent success of deep neural networks depends on the large quantity of high-quality labeled training data. However, the construction of these training data is both labor-intensive and error-prone. For example, photos in the ImageNet dataset [60] were labeled manually by the workers of Amazon Mechanical Turk [36]. These workers need to be trained to follow the pre-defined taxonomy of WordNet, because many of the concepts are unfamiliar to non-experts. Additionally, a well-trained network by a specific dataset may not be an effective model for other datasets or other problems. For instance, as shown in Figure 5–1, given extracted frames from a web video, the state-of-the-art pre-trained deep neural networks of the Google Cloud Platform [46] generate correct labels, but they are either too general or too specific , thus less useful in the annotation. Generating high-quality labeled training set for every task

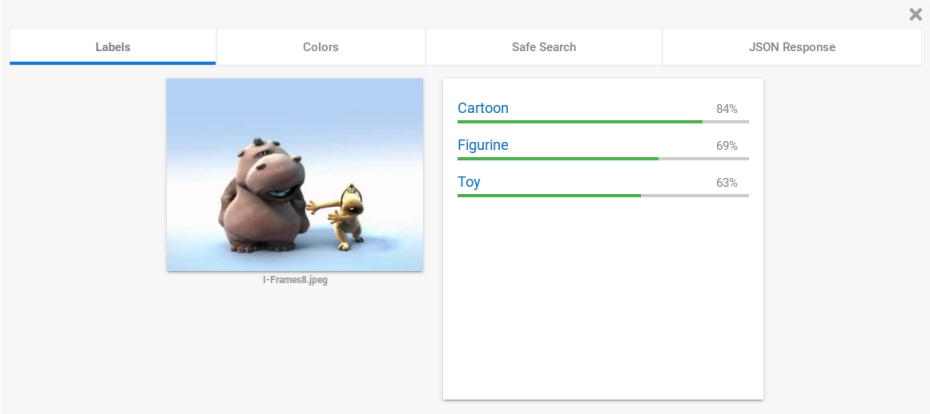


Figure 5–1: Annotating a Keyframe from a Web Video on the Google Cloud Platform

or problem could be costly. Moreover, as of now, deep neural networks are more favorable for image data than for video data, because the video data is much more complicated. The 3-dimensional convolutional neural network (3D CNN) has been applied to video classification [72] and the human action recognition [65]. A fixed number of frames is feed into the 3D network in each training iteration. The training is less robust in capturing the temporal correlations among frames than it does in capturing the spatial correlations within an image.

Multimedia social networking platforms such as Flickr or YouTube allow users to attach description or tags to images/videos. Thus the descriptions and tags can be used as the training labels. However, the dataset obtained from these multimedia social networks are usually sparsely and casually labeled. Besides, lots of users tend to use attractive but irrelevant tags or descriptions with their images/videos to gain a higher popularity. Hence, even with the rich resources of social networks, it is highly difficult to obtain a large training set (especially videos) with high-quality labels for deep neural networks.

There exists a "chicken or the egg" dilemma in this context. The generation of labeled video training set and sanity check of labels rely on the

learning-based models, while the performance of learning relies on the quality and quantity of labeled videos in training sets. To address the dilemma, we develop a graphical model to study the relevance between visual content of multimedia data and their labels, which is capable of starting from a small amount of labeled data and tolerating misleading labels. Next, we explore the similarity relations among multimedia items to expand the graphical model. This design is inspired by the way that humans process unfamiliar observations (i.e., unlabeled data), where people relate the unfamiliar to those that they are familiar with (i.e., labeled data). The use of a graphical model downplays the dependency on high-quality labels by exploiting the joint distribution between labeled and unlabeled data. However, achieving decent estimation accuracy with a graphical model on real-world video data is challenging, mainly due to the following reasons:

- **Limitations of visual features**: For decades, researchers have developed visual features that are handcrafted or learned by neural networks. Every feature captures certain characteristics of multimedia data and loses others. A universal feature set is yet to be discovered, and may not exist at all.
- **Unauthenticated nature of textual content**: Compared with the taxonomy of WordNet used in the ImageNet datasets, the vocabulary of the associated text of online multimedia content is not validated or authenticated by any authority, thus tends to be more personal, ambiguous, noisy, and vulnerable to spamming.

To address these challenges, we devise the graphical model as a multi-view multi-label model. In this model, each view applies a specific similarity metric that quantifies either the visual-visual or label-label distance. To exploit the complementariness and consistency among these views, which is referred to as

learning the "similarity of similarities", we represent these relations as observation functions in Condition Random Field (CRF). Conventionally, CRF is used to encode the known relationships between observations and construct consistent interpretations. One of its advantages is the convenience of representing the higher order relations among the observations. Thus it is powerful to capture the subtle consistency and complementariness between views. Meanwhile, as CRF is complicated in general, we apply the Markov property, which only considers the influence from the neighbors of an item to reduce the computational complexity. Compared with other belief propagation or neighbor voting graphical models [81, 51, 66], ours is capable of describing multiple, higher-order relations (e.g., the visual-visual-label or label-label-visual relations).

The key contributions of this chapter are highlighted as follows:

- We design a graphical model to estimate the relevance of the visual content of multimedia data with associated labels, whose semi-supervised nature reduces its dependency on high-quality labels which are difficult to obtain.
- Based on CRF, we design a multi-view graphical model, which learns the similarity of different similarity metrics defined in different feature spaces (i.e., views), and achieves a decent accuracy of estimation with less advanced visual features.
- Compared with other multi-view or multi-label feature fusion models, our model is a mid-level model. The similarity relations computed with different features remain independent, and the combination weights vary with the change of views or labels.

5.2 Pre-processing: construction of similarity graphs

The cross-modal automatic annotation models are primarily based on the homophily assumption, which implies that visually similar multimedia items

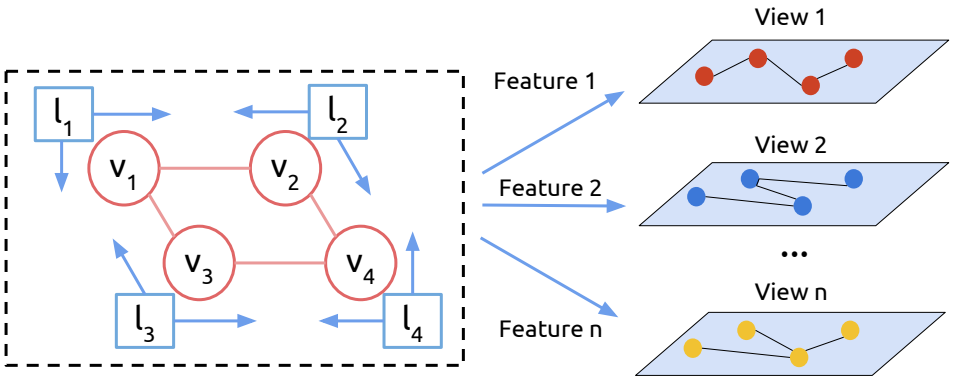


Figure 5–2: Graphs of Similarity Relations. Nodes v_1 to v_4 are videos, and nodes l_1 to l_4 are corresponding labels. Under different visual feature, the similarity relation is different.

can be described by keywords or tags that are semantically similar, and vice versa. Therefore, the relevance estimation commences with the construction of similarity relations on the visual and textual content of the multimedia data.

The similarity relations can be expressed as a graph. As depicted in the left part of Figure 5–2, there is an edge between two images/videos (i.e., v_1 to v_4) if their visual content is similar. In the label propagation approaches, the labels of the images/videos propagate along these edges to other similar items in a weighted way. When the propagation converges, for each image/video, the labels with the highest weights will be regarded as relevant.

Under different visual features, the similarity relations between the images/videos can be different, as shown in the right part of Figure 5–2. The consistency and differences among these relations are helpful information in improving the estimation. Hence a natural extension of the conventional label propagation methods is to construct a composite graph that integrates the similarity relations from various views of the visual features. The examples include hypergraph [38, 138], heterogeneous graph [59, 102, 31], and $\mathcal{K}$-partite [29] (e.g., bipartite [51]) graph. The objective of these approaches is to fuse

the multiple types of similarity graphs into a general similarity graph, upon which the probability of propagating the labels to relevant videos increases.

In this thesis, we postulate that the labels are connected to the visual content of multimedia data in a variety of manners, depending on which visual feature is extracted. Therefore, we maintain the similarity graphs constructed under different features to be independent to incorporate the diversity about how the visual and textual modalities are related to each other.

Besides, the semantic similarities between the labels are absent in the model shown in the left part of Figure 5–2. In fact, each multimedia item has multiple labels, and the probability of co-occurrence of specific labels is significantly higher than with other combinations. Depicting the connections between the visual similarity graphs, and their relationships with the semantic correlations of the labels would generate a highly convoluted graph. Therefore, in this thesis, the semantic correlations among the labels are regarded as another similarity relations view.

Calculating the similarities between every pair of multimedia items is exhaustive when the volume of multimedia data grows to the web scale. Moreover, according to the homophily assumption, visually different multimedia items tend to have semantically different labels. Therefore, in the construction of the similarity graph, we add an edge between items only if they are adjacent to each other in particular visual feature space. Specifically, we apply the LSH (Locality Sensitive Hashing) technique [6] to determine whether a pair of multimedia items are neighbors. In this way, not only the cost of both the similarity graph constructions is reduced, but also the calculations of the CRF model are accelerated.

When the pre-processing phase completes, for each multimedia item v, its nearest neighboring items v's in the space of each visual feature f, along with

the distances between the items in the feature spaces $d_v(v, v', f)$ obtained. In addition, for each label l, the most semantically correlated labels l' in the vocabulary are also retrieved, and their distances $d_l(l, l')$ are calculated.

5.3 The multi-view multi-label graphical model

The estimation of the cross-modal relevance, or the relevance between the visual content and the labels, can be formally defined as the estimation of the conditional probability $p(l_i = val|v)$, where v refers to a multimedia item, $l_i, i = 0, \cdots, n$ is the i-th candidate label for annotation, and $val = 0, 1$ is the value that indicates whether l_i is a proper label for v.

In this section, we propose a probabilistic graphical model to predict this probability. This model is based on CRF and exploits the consistency and differences between multiple similarity relations on both the visual and textual modalities.

5.3.1 Graphical structure design

According to the homophily assumption, visually similar items share semantically correlated textual descriptions. Therefore, the similarity relations obtained at the preprocessing phase on both the visual and textual modalities can be incorporated in the cross-modal relevance estimation.

Similarity relations as observation functions

In this thesis, we divide the cross-modal connections between the visual content and the presence/absence of a label into three categories:

1. A label is present or absent in the textual description of an image/video provided by Internet users;
2. Suppose the labels of multimedia data are known, and the visual content of multimedia data is hidden except for the similarity relations on this modality. According to the homophily assumption, we postulate that the presence/absence of a label in the description of a multimedia item

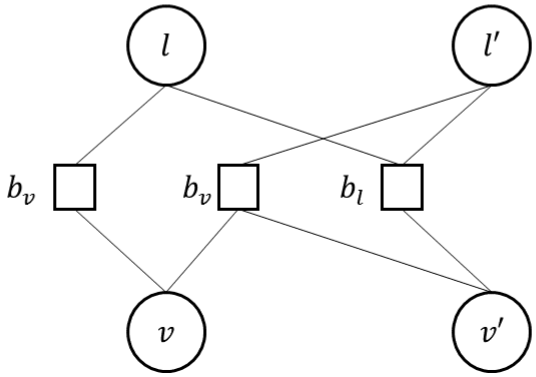


Figure 5–3: Three Types of Observation Functions

can be inferred from the presence/absence of the label in the description of the neighboring items;

3. By switching the observability of the labels and visual content, the other perspective of the homophily assumption can be applied. More specifically, suppose the visual content is observable, and the labels are hidden except for their semantic correlations. For a given image/video, we assume that if a label is present or absent in its description, the probabilities of the occurrence/absence of the semantically correlated labels are significantly higher.

Through the three cross-modal connections, the similarity relations obtained by applying the LSH technique can be incorporated in the relevance estimation. To integrate these connections into a CRF model, we design three types of observation functions. For a given v, if l is present in the description, $val = 1$; otherwise, $val = 0$.

1. The first type of observation function $b_v(v, l, val)$ is defined as:

$$b_v(v, l, val) = p(l = val|v) \tag{5.1}$$

.

2. Suppose an image/video in the visual neighborhood of v is denoted as v', and the presence/absence of a label l' in the description of v' is established. This type of function can be defined as:

$$b_v(v, v', l', val) = \exp\{-\alpha d_v^2(v, v')\} \cdot p(l = val|v'), v' \in U(v) \tag{5.2}$$

where α is a parameter that controls the influence of distance between v and v' in the model.

3. The third type of observation function indicates that if label l' is semantically similar to l, then l' is probably a suitable label for interpreting v. It is defined as:

$$b_l(l, l', v, val) = \exp\{-\beta d_l^2(l, l')\} \cdot p(l' = val|v), l' \in U(l) \tag{5.3}$$

where β is a parameter that has a similar function as α.

The first and second type of observation functions can be merged because an image/video itself is in its visual neighborhood. Thus both of them are represented by the b_v observation functions. The second and third kinds of relations are more complicated, high-order relations than the first kind. Furthermore, to the right side of Eq. 5.1, 5.2, and 5.3, the term $p(l = val|v)$ appears in different forms. This term is the objective of our model. Consequently, our model is an iterative model. In the beginning, the value of $p(l = val|v)$ is determined by the original label assignments from users. More specifically, if users assign a label l to a multimedia item v, then $p(l = 1|v) = 1$. Otherwise, $p(l = 0|v) = 1$.

The mid-level fusion of multiple views

Representing the three kinds of relations as observation functions makes the incorporation of multiple views (i.e., similarity relations) smooth because these functions are treated as factors in the CRF model coherently.

For the representation v^{f_i} of v under each feature $f_i, i = 1, \ldots, n$, the first and second observation functions, or the functions on the visual modality, can be extended to $b_v(v, v', l', val, f_i)$, which is defined as,

$$b_v(v, v', l', val, f_i) = \exp\{-\alpha d_v^2(v, v', f_i)\} \cdot p(l = val|v'), v' \in U_{f_i}(v). \tag{5.4}$$

Integrating multiple features and extending the CRF model to be a multi-view model substantially increase the number of all the three types of observation functions, as shown in Figure 5–4:

1. For the representation v^{f_i} of v under each feature $f_i, i = 1, \ldots, n$, the corresponding functions $b_v(v, v, l, val, f_i)$ will be added to the model;
2. In the view induced from a visual feature f_i, if v' is in the neighborhood of v, $U_{f_i}(v)$, observation functions $b_v(v, v', l', val, f_i)$ that describe the high-order relations among v, v', and l', are introduced to the model. Under different features, the neighboring relations of the same pair of multimedia items can be different. For example, we may both have $v' \in U_{f_i}(v)$ and $v' \notin U_{fj}(v), j \neq i$;
3. Akin to the first point, when a label l is connected to v' through l', for each v^{f_i}, $i = 1, \ldots, n$, the corresponding $b_l(l, l', v, val)$ functions will be added.

With these observation functions, the multi-view CRF model is defined as,

$$p(l = val|v) = \frac{1}{Z(v, l)} \exp\{\sum_{v'} \lambda_v(v', l', val, f_i) \cdot b_v(v, v', l', val, f_i) + \sum_{l'} \lambda_l(l', val) \cdot b_l(l, l', v, val)\}, \tag{5.5}$$

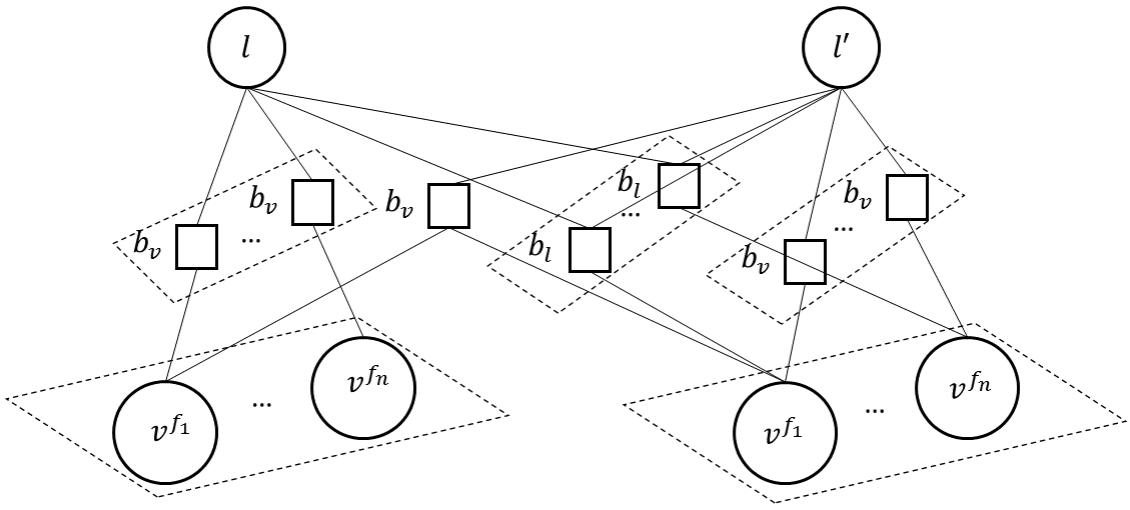


Figure 5–4: The Structure of the Multi-View Graphical Model.

where λ_v and λ_l are parameters to be learned, and $Z(v, l)$ is the normalizing factor over the possible values of the labels,

$$Z(v,l) = \sum_{val=0}^{1} \exp\{\sum_{v'} \lambda_v(v', l', val, f_i) \cdot b_v(v, v', l', val, f_i) + \sum_{l'} \lambda_l(l', val) \cdot b_l(l, l', v, val)\}. \quad (5.6)$$

In contrast to the feature-level fusion model where a new feature space is learned from various feature spaces or the decision-level model which polls multiple independent learners, the fusion of the multiple visual and semantic similarity relations by a CRF model is a mid-level integration. The similarity relations computed with different visual features remain independent in our model, yet their combination can be varied by adjusting λ_v and λ_l.

The combination of views of our model is fine-grained, as the weights λ_v and λ_l are different for each image/video and each label. Concretely, λ_vs are indexed by (v, l, val, f_i), which implies that from the perspective of the visual modality, the association of a label to a multimedia item is distinctive in each visual feature space. Additionally, λ_ls are by (l, val), which suggests that the

correlation of one label to another label varies with the presence/absence of the latter one.

5.3.2 Learning

The input to the model in the learning process is the feature representations of the images/videos in the training set, and the relevance score (i.e., 0 or 1) of these images/videos to the pre-defined semantic labels. For the images/videos in the test set, the values of the b_v and b_l functions can be calculated similarly, except that the neighborhood of a test image/video is comprised of the images/videos in the training set.

The estimation of model parameters

The parameters of the model are calculated by using maximum likelihood estimation. Given the training set

$$\begin{aligned} D = \{&(v_1, l_1, val(v_1, l_1)), (v_1, l_2, val(v_1, l_2)), \ldots, \\ &(v_1, l_n, val(v_1, l_n)), (v_2, y_2, val(v_2, y_2)), \ldots, \\ &(v_m, y_n, val(v_m, y_n))\}, \end{aligned}$$

the log likelihood function is,

$$\begin{aligned} L(\Lambda|D) &= \log(\prod_{d=1}^{|D|} p(l_d = val_d|v_d)) - \sum_{v,l,val,f_i} \frac{\lambda_v^2(v,l,val,f_i)}{2\sigma^2} - \sum_{l,val} \frac{\lambda_l^2(l,val)}{2\sigma^2} \\ &= \sum_{d=1}^{|D|} \{\sum_{v'} \lambda_v(v', l_d, val_d, f_i) b_v(v_d, v', l_d, val_d, f_i) + \sum_{l'} \lambda_l(l', val_d) b_l(l_d, l', v_d, val_d) - \log Z(v_d, l_d)\} \\ &\quad - \sum_{v,l,val,f_i} \frac{\lambda_v^2(v,l,val,f_i)}{2\sigma^2} - \sum_{l,val} \frac{\lambda_l^2(l,val)}{2\sigma^2}, \quad (5.7) \end{aligned}$$

where the last three terms are Gaussian prior used to reduce over-fitting.

The partial derivatives of $L(\Lambda|D)$ with respect to parameters λ_v are,

$$\frac{\partial L(\Lambda|D)}{\partial \lambda_v(v', l_d, val_d, f_i)} = \sum_{d=1}^{|D|} \{b_v(v_d, v', l_d, val_d, f_i) - \frac{1}{Z(v_d, l_d)} \frac{\partial Z(v_d, l_d)}{\partial \lambda_v(v', l_d, val_d, f_i)}\} - \sum_{v,l,val,f_i} \frac{\lambda_v(v, l_d, val_d, f_i)}{\sigma^2}, \tag{5.8}$$

and $\frac{\partial}{\partial \lambda_v(v', l_d, val_d, f_i)} Z(v_d, l_d)$ can be calculated as,

$$\begin{aligned}
\frac{\partial Z(v_d, l_d)}{\partial \lambda_v(v', l_d, val_d, f_i)} &= \frac{\partial}{\partial \lambda_v(v', l_d, val_d, f_i)} \sum_{val=0}^{1} \exp\{\sum_{v'} \lambda_v(v', l_d, val_d, f_i) b_v(v_d, v', l_d, val_d, f_i) \\
&\quad + \sum_{l'} \lambda_l(l', val_d) b_l(l_d, l', v_d, val_d)\} \\
&= \sum_{val_d=0}^{1} \frac{\partial}{\partial \lambda_v(v', l_d, val_d, f_i)} \exp\{\sum_{v'} \lambda_v(v', l_d, val_d, f_i) b_v(v_d, v', l_d, val_d, f_i) \\
&\quad + \sum_{l'} \lambda_l(l', val_d) b_l(l_d, l', v_d, val_d)\} \\
&= \exp\{\sum_{v'} \lambda_v(v', l_d, val_d, f_i) b_v(v_d, v', l_d, val_d, f_i) + \sum_{l'} \lambda_l(l', val_d) b_l(l_d, l', v_d, val_d)\} \\
&\quad \cdot b_v(v_d, v', l_d, val_d, f_i).
\end{aligned} \tag{5.9}$$

Thus $\frac{1}{Z(v_d, l_d)} \frac{\partial}{\partial \lambda_v(v', l_d, val_d, f_i)} Z(v_d, l_d)$ is,

$$\begin{aligned}
\frac{1}{Z(v_d, l_d)} \frac{\partial Z(v_d, l_d)}{\partial \lambda_v(v', l_d, val_d, f_i)} &= b_v(v_d, v', l_d, val_d, f_i) \frac{1}{Z(v_d, l_d)} \exp\{\sum_{v'} \lambda_v(v', l_d, val_d, f_i) \\
&\quad b_v(v_d, v', l_d, val_d, f_i) + \sum_{l'} \lambda_l(l', val_d) b_l(l_d, l', v_d, val_d)\} \\
&= p(l_d = val_d|v_d) b_v(v_d, v', l_d, val_d, f_i).
\end{aligned} \tag{5.10}$$

Plug Eq. 5.10 into Eq. 5.8 we get,

$$\frac{\partial L(\Lambda|D)}{\partial \lambda_v(v', l_d, val_d, f_i)} = \sum_{d=1}^{|D|} \{b_v(v_d, v', l_d, val_d, f_i)(1 - p(l_d = val_d|v_d))\} - \sum_{v,l,val,f_i} \frac{\lambda_v(v', l_d, val_d, f_i)}{\sigma^2}. \tag{5.11}$$

Similarly, we have,

$$\frac{\partial L(\Lambda|D)}{\partial \lambda_l(l', val_d)} = \sum_{d=1}^{|D|} \{b_l(l_d, l', v_d, val_d)(1 - p(l_d = val_d|v_d))\} - \sum_{l,val} \frac{\lambda_l(l', val_d)}{\sigma^2}. \tag{5.12}$$

Since these equations cannot be solved in a closed form, we use a fast convex optimization method, the L-BFGS (Limited-memory Broyden–Fletcher–Goldfarb–Shanno) algorithm to find the global maximum of the likelihood function $L(\Lambda|D)$.

5.4 Evaluation

In this section, we evaluate our cross-modal relevance estimation model against other quintessential feature fusion models from the literature, Multiple Kernel Learning (MKL) and Ensemble Learning (EL). The MKL model also belongs to the category of mid-level feature fusion models, and the EL model is a decision-level fusion model. The feature-level models, such as deep neural networks, can be naturally integrated into the mid-level and decision-level models. Hence they are excluded from the benchmark models.

5.4.1 Experiment settings

All the experiments are conducted on a commodity computer, which is equipped with an 8-core Intel i7-3632 CPU, 8GB memory, and Ubuntu 16.04 system. The programs are coded in Python, with the support of the libraries of Numpy, SciPy 0.18, and scikit-learn 0.18.2.

Dataset. We adopt the Columbia Consumer Video (CCV) database [67] instead of the CC_WEB_VIDEO dataset to evaluate our feature fusion model against other benchmark models because the ground-truth about the relevance between videos and semantic concepts are provided along with this dataset. The CCV database contains 9317 web videos and is manually annotated through Amazon MTurk with 20 semantic concepts. In these videos, 4659 are used for training and 4658 for evaluations.

Three types of features are extracted and represented as multi-dimensional vectors from the videos in the CCV database. The textual concepts are represented as vectors as well. These feature vectors can be briefly described as follows:

- MFCC: The MFCC (Mel-frequency Cepstral Coefficients) is a widely used audio feature. Through the audio processing, the short-term power spectrum of the sound of the videos is represented as 5000-dimensional vectors.
- SIFT: The SIFT (Scale-invariant feature transform) has been employed in a wide range of visual recognition and categorization applications. The interested points in the frames of videos are located and clustered into visual words. The dimension of this type of representation is 5000.
- STIP: The STIP (Spatial-temporal interest points) captures not only the spatial distribution of the visual words as the SIFT feature but also the temporal flow of these points. The dimension of this type of feature vectors is also 5000.
- Word Representation: The semantic concepts are represented as vectors of 300 dimensions in the database of GloVe [100]. The GloVe word representation captures the co-occurrence information of words.

With these vectors, different views (i.e., similarity relations) can be established rapidly with approximate hashing methods (e.g., LSH). To overcome the curse-of-dimensionality issue, we adopt the FALCONN library which has solid performance for the high-dimensional vectors. When the identification of neighboring videos or semantic concepts completes, the Euclidean distances between a video/concept and its neighbors are calculated.

Benchmark models. The following feature fusion models only combine the visual and audio features and neglect the similarity relations of the

semantic concepts. Both models are kernel-based. For each concept, the kernel is trained to fit the relevance/irrelevance of this concept to the videos in the training set. The weights for the combination of the kernels or the predictions from the kernels are regarded as hyperparameters. Through grid search, the weights that achieve the highest Average Precision score on the test set are designated for the combination.

- Multiple Kernel Learning (MKL): The essence of the MKL algorithm that appears in literature is the same as [39]. Following the descriptions, we construct multiple linear kernels and combine them with the weights obtained as mentioned above. The combined kernel is trained to fit each concept by using the SVM model.
- Ensemble Learning (EL): In this fusion model, we train multiple SVM models for each concept. The predictions of these models are combined with the weights obtained from the grid search.

Evaluation metric. In this thesis, the performance of the cross-modal relevance estimation is evaluated by the Mean Average Precision (MAP) of the classification task. The Mean Average Precision is the average of the Average Precision (AP) of the classification of the videos over the 20 semantic concepts, which is calculated as:

$$AP(l) = \frac{1}{R} \sum_{i=1}^{n} \frac{r_i}{i} \delta(v_i, l), \tag{5.13}$$

where R is the number of relevant videos for a given concept l and r_i is the number of relevant videos in the top i ranked videos. The function $\delta(\cdot)$ is an indicator function, which equals to 1 if v_i and l are relevant, and 0 otherwise.

5.4.2 Experimental results

We implemented our CRF-based feature fusion model, along with the MKL and EL models. All of these models are trained and tested on the same

Table 5–1: Mean Average Precision Comparison

MODEL	MAP (%)
MKL	45.60
EL	37.83
Ours	51.53

training and test sets. For all the experiments in this section, we empirically set the parameters of our model as $\alpha = 5$, and $\beta = 5$. Before the training, all the λs are initialized to 1.0, to ensure that the effect of every observation function is taken into account.

Average precision

The optimal parameters for the three models are set by running the models on the test set and tracking the variation of the average precision values. The highest MAP scores of the three models are reported in Table 5–1. As shown in this table, these results imply that in the prediction of the three most probable concepts for a video, the three models make at least one correct answer. Also, our CRF-based model outperforms the MKL and EL feature fusion models.

The comparison can be conducted over the 20 semantic labels. From Figure 5–5, the mid-level models (i.e., MKL, CRF) consistently outperform the decision-level model (i.e., EL), which demonstrates the advantage of conducting the feature fusion before predictions. Additionally, compared with the MKL model, the weight parameters for feature fusion of our model are distinctive not only for different types of features but also for different labels. In another word, we assume that the relevancy between a label and a video represented by a visual feature varies with both the label and the visual feature. Consequently, our model surpasses the MKL model on more semantic concepts.

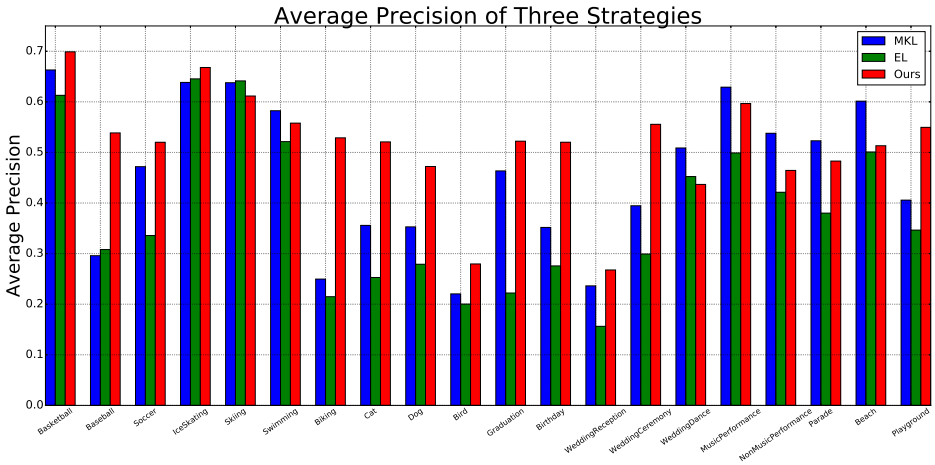


Figure 5–5: The Comparison of the Average Precision of the three Strategies

Relevant factors

The performance of our model can be affected by a variety of factors. Among these factors, whether the model is trained separately for each label, the number of neighboring videos, and the regularization terms of Eq. 5.7 are most relevant. Their effects on the Average Precision of our model for each label are investigated in this part.

Label-wise training. According to Eq. 5.5, our model can be trained and tested for a specific label, or all the labels in the database. The former training style is referred to as the label-wise training, and the latter one is called all-labels training in this thesis. In Figure 5–6, all-labels training outperforms label-wise training for each label. It follows that there are connections between labels that can only be captured by the all-labels training rather than the label-wise training.

The number of neighbors. Intuitively, the more powerful a feature is, the more neighboring videos should be included in the computations of our model. However, since we have no prior knowledge about the effectiveness of the visual and audio features, the quantity of the neighbors (including the

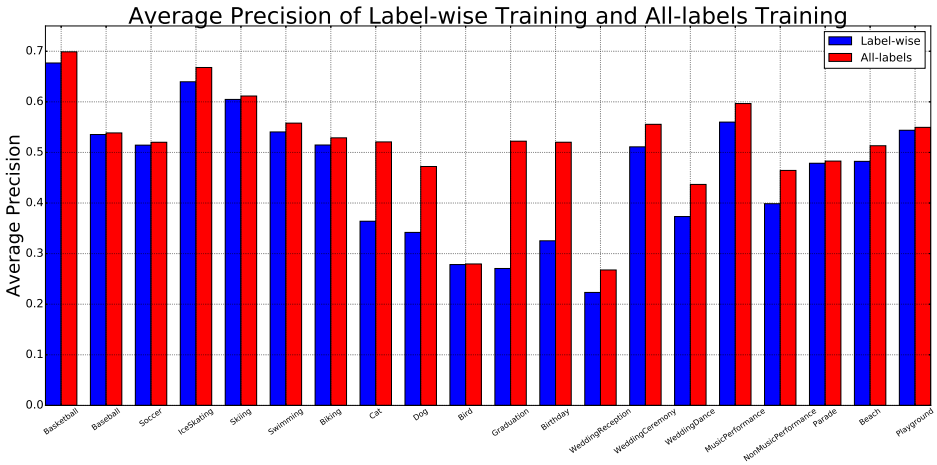


Figure 5–6: The Comparison of the Average Precision of the Label-wise Training and All-labels Training

video itself) of a video is set to the same value for the three types of features in the pre-processing stage. In the experiments, the number of neighbors varies from 3 to 11. As for the GloVe word representations, the quantity of the neighbors is set to 3 because there are only 20 semantic concepts in the database.

From Table 5–2 ($\sigma = 2.0$), we observe that:

- More neighbors do not guarantee higher average precision score.
- The optimal value for the number of neighboring videos is distinctive for each label.
- Generally speaking, if the model is more effective for the predictions of a label than another label, its optimal setting for the number of neighbors is higher.

Regularization. In this thesis, we adopt the Gaussian prior to regularize the training of λs. For the experiments in this part, we set the quantity of the neighbors to 5, and vary σ from 2.0 to 5.0. The scores of the average precision of the model are reported in Figure 5–3.

Table 5–2: The Effect of the Quantity of Neighbors on AP

LABEL	3Nhs (%)	5Nhs (%)	7Nhs (%)	9Nhs (%)	11Nhs (%)
Basketball	61.48	57.50	49.92	66.17	42.82
Baseball	25.27	33.64	24.46	13.43	16.36
Soccer	34.85	32.17	32.99	33.42	41.70
IceSkating	63.51	44.59	45.53	45.50	35.01
Skiing	53.90	54.48	53.41	51.88	24.93
Swimming	50.27	49.97	48.10	48.73	16.89
Biking	19.55	27.18	18.40	51.47	1.47
Cat	24.55	21.41	25.42	14.83	7.41
Dog	25.58	26.57	22.47	17.50	16.66
Bird	21.76	14.37	17.63	7.63	8.22
Graduation	24.13	18.02	14.95	12.28	13.52
Birthday	12.49	13.20	19.23	18.69	1.72
WeddingReception	11.97	12.89	22.33	11.23	11.77
WeddingCeremony	45.24	51.11	24.03	40.01	27.59
WeddingDance	29.88	26.75	32.14	24.87	30.10
MusicPerformance	49.67	50.54	41.05	44.02	41.49
NonMusicPerformance	37.24	38.56	39.87	29.82	26.70
Parade	44.22	45.60	44.26	33.43	37.01
Beach	41.60	39.58	27.51	40.66	40.97
Playground	27.41	29.99	33.90	54.39	12.38

Table 5–3: The Effect of the Regularization on AP

LABEL	$\sigma = 2.0$ (%)	$\sigma = 3.0$ (%)	$\sigma = 4.0$ (%)	$\sigma = 5.0$ (%)
Basketball	57.50	64.11	63.78	64.83
Baseball	33.64	25.24	22.17	25.05
Soccer	32.17	42.67	43.19	39.29
IceSkating	44.59	58.22	59.67	62.17
Skiing	54.48	56.40	56.81	57.13
Swimming	49.97	53.02	52.65	51.29
Biking	27.18	18.85	13.29	10.16
Cat	21.41	36.40	28.81	23.56
Dog	26.57	25.38	25.33	20.87
Bird	14.37	12.73	15.73	13.85
Graduation	18.02	17.80	18.20	15.09
Birthday	13.20	18.62	23.21	26.35
WeddingReception	12.89	12.99	10.46	11.05
WeddingCeremony	51.11	49.06	45.63	46.95
WeddingDance	26.75	34.58	31.50	32.62
MusicPerformance	50.54	56.00	50.15	52.86
NonMusicPerformance	38.56	37.86	36.60	35.76
Parade	45.60	44.75	47.13	47.86
Beach	39.58	40.38	45.75	42.33
Playground	29.99	27.32	27.56	27.29

Akin to the impacts of the number of neighbors, the optimal value of σ is also different from label to label. For specific labels such as "IceSkating", applying regularization enhances the score of average precision (to 62.17), which is normally around 40. This impact reminds us that choosing a proper regularization prior can be vital to the accuracy of our model.

CHAPTER 6
Conclusions and Future Work

In this chapter, we conclude the thesis. The contributions presented in the thesis are summarized in Section 6.1. The challenges and potential research directions are outlined in Section 6.2

6.1 Conclusions

Due to the intrinsic complexity and the unprecedentedly vast volume, the automatic annotation of the online multimedia data is much more challenging than other types of data. In this thesis, we attempted to approach this problem by the homophily assumption, through which the intra-modal similarity relations of the visual and textual modalities of the multimedia data are utilized to estimate the relevance of the two modalities. This assumption implies the roadmap that we can follow to resolve the annotation problem. The first step of the roadmap is to identify the similarity relations between the multimedia items under visual and textual feature, which corresponds to the first and a part of the second project in this thesis. The second step is to infer the cross-modal relevance with these similarity relations, which corresponds to the third project and a part of the second project.

Similarity detection of the online videos

In the first project, we proposed and developed an efficient NDVD cloud system, called CompoundEyes, by using a new detection paradigm. Instead of designing a sophisticated video representation, the focus has been shifted to the design of a well-organized system. Rather than feature design, we introduced improvements in accuracy through classifiers. Because of reduced dimensionality and parallelism, we reduced the duration required for precise

duplicate detection. Moreover, experiments and analysis corroborated that CompoundEyes not only outperforms contemporary NDVD and NDVR systems, but also the feature-centered systems based on BoWs and CNN features regarding accuracy. At the same time, CompoundEyes bested or matched its peers in both peak memory usage and time complexity. In conclusion, CompoundEyes is feasible and practical to perform large-scale NDVD tasks in the cloud. As other NDVD/NDVR systems, CompoundEyes needs a training set and ground-truth labels, the acquisition of which is beyond the scope of this thesis.

Empirical Study of the cross-modal relevance of the online videos

In the second project, with CompoundEyes, we conducted an empirical study of the properties of textual content on a popular real-world online video dataset, CC_WEB_VIDEO. The following insights can be drawn from the analysis of this dataset:

- **Textual Content as Complex System**: After the removal of non-word parts and function words, the word count distribution displays a typical Zipf's Law feature, which implies that this dataset is sufficiently large and complex, and there may exist a hierarchy in the vocabulary, and social structures of users.
- **The Sparsity of the Textual Content**: Other than titles, the number of words used to describe the visual content is less than 40 for most of the videos in the dataset. With descriptions of this length, it is not expected that the visual content can be semantically illustrated. Performing similarity comparison on the textual content would be unstable and less satisfactory compared with using the visual content.
- **The Poor Quality of the Textual Content**: Though affected by various factors, the poor quality of the textual content is one of them

which reduces the relevance rate of search results. More keyword matches do not guarantee more reliable results.

- **URL's Occurrence is not a Satisfactory Heuristic Spamming Indicator**: By data analysis, we showed that the occurrence of URL is not tightly related to video spamming.
- **The Coherence between the Visual and Textual Content**: Similar visual content does accompany similar textual content and vice versa. Although with more degree of freedom of word choice, this similarity is less stable than using visual content.

In the CC_WEB_VIDEO dataset, the 24 search keywords are also the 24 topic names. If all the videos that are searched with the same keyword perfectly matched the topic, both the visual and textual similarity experiments would be dominated by values close to 1. In contrast, the low values of similarity scores suggest that 24 topics are not sufficient.

With latent models and sub-category approaches, we can keep refining the topics. However, the experiment results in the coherence between visual and textual contents entail that the complexity of their relationships is beyond the capacity of these approaches. We found that the textual similarity scores can be affected by various factors, such as the length of the piece of text, textual content field, and textual similarity measures themselves. Textual content is not merely noisy, but logically ambiguous and misguided. Two videos that have similar titles may be entirely different in tags and descriptions, and the opposite can also be true. When all of these factors are taken into account, the relationships between them, and with the visual content, become chaotic.

These insights about the textual content of online videos suggest that combining the visual and textual content directly, as tempting as it appears to bridge the semantic gap between low-level visual features and high-level

semantic meanings, is problematic. The sparsity, poor quality, and the high degree of freedom of expressiveness of textual content are impediments for any model to achieve satisfactory performance.

Cross-modal relevance estimation

In this project, based on CRF, we developed a probabilistic, multi-view multi-label graphical model to estimate the relevance of the visual and textual modalities of online videos. This model takes the original video-label associations, linguistic label similarities, and visual similarities into account to deal with the complicated relationships between the two modalities of videos. It is designed as a multi-view model to overcome the limitations of individual visual features in identifying visually adjacent videos. Meanwhile, the architecture of this model is open, which means that whenever a more effective new feature (visual or texture) is discovered, it can be incorporated into the model seamlessly. We implemented this model and compared it with two other feature fusion models. Through the experiments, we demonstrated that our model consistently outperforms the benchmark models.

6.2 Future Work

The automatic annotation of the multimedia data is an important and challenging research topic. It is also an essential step towards the understanding of the content of the multimedia data. There are numerous attractive and influential problems in this area. By following the roadmap planned in this thesis, we mainly focus on the issues that belong to the following categories:

Visual features for videos. In the simple features combined by CompoundEyes, only the motion orientation feature captures the temporal information between consecutive frames. Hence CompoundEyes is ineffective for the more complicated temporal variations of videos. However, due to the sequential nature of the temporal dimension, the fusion of the spatial and

temporal information makes the processing of videos less parallelable. We intend to investigate the temporal variations of videos, dissect the temporal dimension into multiple independent segments, and integrate these temporal segments with corresponding spatial areas.

Automatic annotation of the multimedia data. The graphical model developed in this thesis can also be applied to image datasets, as long as the visual features for images are available. Employing powerful features, such as the ones discovered by deep neural networks, is expected to enhance the accuracy of cross-model relevance estimation considerably. Also, the labels provided by the users of social networks tend to be subjective and might be semantically distant from the predictions of the current object/scenery classification models. It is challenging to connect the subjective labels and objective classification results.

Distributed computing. Currently, distributed computing techniques and platforms such as Spark have been becoming an integral part of the processing techniques of the online multimedia data, due to their sheer volume and unprecedentedly growing speed. In the first project, the feature extraction algorithms and the architecture of CompoundEyes are designed to be intrinsically distributable, thus migrating the system to the state-of-the-art distributed computing platforms is smooth. Compared with the shared-memory architecture applied in the project, a real distributed computing platform can further boost the throughput of the similarity detection/retrieval system. Moreover, the multi-view multi-label graphical model proposed in the third project can also be parallelized on Spark, because the computations of λs can be interpreted as map-reduce style key/value aggregations.

References

[1] 5 advanced youtube seo tactics to drive more traffic to your videos and website, 2014.

[2] Facebook app can answer basic questions about whats in photos, 2015.

[3] Spam, deceptive practices, and scams, 2015.

[4] Palakorn Achananuparp, Xiaohua Hu, and Xiajiong Shen. The evaluation of sentence similarity measures. In *Data warehousing and knowledge discovery*, pages 305–316. Springer, 2008.

[5] Arnon Amir, Marco Berg, Shih-Fu Chang, Winston Hsu, Giridharan Iyengar, Ching-Yung Lin, Milind Naphade, Apostol Natsev, Chalapathy Neti, Harriet Nock, et al. Ibm research trecvid-2003 video retrieval system. *NIST TRECVID-2003*, 2003.

[6] Alexandr Andoni, Piotr Indyk, Thijs Laarhoven, Ilya Razenshteyn, and Ludwig Schmidt. Practical and optimal lsh for angular distance. In *Advances in Neural Information Processing Systems*, pages 1225–1233, 2015.

[7] Melanie Aurnhammer, Peter Hanappe, and Luc Steels. Integrating collaborative tagging and emergent semantics for image retrieval. In *Proc. of the Collaborative Web Tagging Workshop (WWW?6)*, 2006.

[8] Jay Baer. 5 snapchat statistics that prove its power.

[9] Song Bai, Shaoyan Sun, Xiang Bai, Zhaoxiang Zhang, and Qi Tian. Smooth neighborhood structure mining on multiple affinity graphs with applications to context-sensitive similarity. In *European Conference on Computer Vision*, pages 592–608. Springer, 2016.

[10] Fabricio Benevenuto, Tiago Rodrigues, Virgílio AF Almeida, Jussara Almeida, Marcos Gonçalves, and Keith Ross. Video pollution on the web. *First Monday*, 15(4), 2010.

[11] Jonathan Bragg, Daniel S Weld, et al. Crowdsourcing multi-label classification for taxonomy creation. In *First AAAI conference on human computation and crowdsourcing*, 2013.

[12] brainHQ. How vision works.

[13] Michael M Bronstein, Alexander M Bronstein, Fabrice Michel, and Nikos Paragios. Data fusion through cross-modality metric learning using similarity-sensitive hashing. In *CVPR*, volume 1, page 5. San Francisco, CA, 2010.

[14] Serhat Selcuk Bucak, Rong Jin, and Anil K Jain. Multi-label learning with incomplete class assignments. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 2801–2808. IEEE, 2011.

[15] Liangliang Cao, Jiebo Luo, and Thomas S Huang. Annotating photo collections by label propagation according to multiple similarity cues. In *Proceedings of the 16th ACM international conference on Multimedia*, pages 121–130. ACM, 2008.

[16] Ciro Cattuto, Dominik Benz, Andreas Hotho, and Gerd Stumme. *Semantic grounding of tag relatedness in social bookmarking systems*. Springer, 2008.

[17] Ciro Cattuto, Vittorio Loreto, and Luciano Pietronero. Semiotic dynamics and collaborative tagging. *Proceedings of the National Academy of Sciences*, 104(5):1461–1464, 2007.

[18] Xiangyu Chen, Yadong Mu, Shuicheng Yan, and Tat-Seng Chua. Efficient large-scale image annotation by probabilistic collaborative multi-label propagation. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 35–44. ACM, 2010.

[19] Yixin Chen, Jinbo Bi, and James Ze Wang. Miles: Multiple-instance learning via embedded instance selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1931–1947, 2006.

[20] Yixin Chen, Wenbo He, Yu Hua, and Wen Wang. Compoundeyes: Near-duplicate detection in large scale online video systems in the cloud. In *Computer Communications, IEEE INFOCOM 2016-The 35th Annual IEEE International Conference on*, pages 1–9. IEEE, 2016.

[21] Ed H Chi and Todd Mytkowicz. Understanding the efficiency of social tagging systems using information theory. In *Proceedings of the nineteenth ACM conference on Hypertext and hypermedia*, pages 81–88. ACM, 2008.

[22] Chih-Yi Chiu, Chu-Song Chen, and Lee-Feng Chien. A framework for handling spatiotemporal variations in video copy detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 18(3):412–417, 2008.

[23] Chih-Yi Chiu, Tsung-Han Tsai, Yu-Cyuan Liou, Guei-Wun Han, and Hung-Shuo Chang. Near-duplicate subsequence matching between the continuous stream and large video dataset. *IEEE Transactions on Multimedia*, 16(7):1952–1962, 2014.

[24] Chien-Li Chou, Hua-Tsung Chen, and Suh-Yin Lee. Pattern-based near-duplicate video retrieval and localization on web-scale videos. *IEEE Transactions on Multimedia*, 17(3):382–395, 2015.

[25] C Christoudias, Raquel Urtasun, and Trevor Darrell. Multi-view learning in the presence of view disagreement. *arXiv preprint arXiv:1206.3242*, 2012.

[26] Lingyang Chu, Shuqiang Jiang, Shuhui Wang, Yanyan Zhang, and Qingming Huang. Robust spatial consistency graph model for partial duplicate image retrieval. *IEEE Transactions on Multimedia*, 15(8):1982–1996, 2013.

[27] Ondřej Chum, James Philbin, Michael Isard, and Andrew Zisserman. Scalable near identical image and shot detection. In *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 549–556. ACM, 2007.

[28] Peng Cui, Zhiyu Wang, and Zhou Su. What videos are similar with you?: Learning a common attributed representation for video recommendation. In *Proceedings of the ACM International Conference on Multimedia*, pages 597–606. ACM, 2014.

[29] Dingxiong Deng, Fan Bai, Yiqi Tang, Shuigeng Zhou, Cyrus Shahabi, and Linhong Zhu. Label propagation on k-partite graphs with heterophily. *arXiv preprint arXiv:1701.06075*, 2017.

[30] Carlotta Domeniconi and Bojun Yan. Nearest neighbor ensemble. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 1, pages 228–231. IEEE, 2004.

[31] Dhivya Eswaran, Stephan Günnemann, Christos Faloutsos, Disha Makhija, and Mohit Kumar. Zoobp: Belief propagation for heterogeneous networks. *Proceedings of the VLDB Endowment*, 10(5), 2017.

[32] Jianping Fan, Yi Shen, Chunlei Yang, and Ning Zhou. Structured max-margin learning for inter-related classifier training and multilabel image annotation. *Image Processing, IEEE Transactions on*, 20(3):837–854, 2011.

[33] Jianping Fan, Yi Shen, Ning Zhou, and Yuli Gao. Harvesting large-scale weakly-tagged image databases from the web. In *CVPR*, volume 10, pages 802–809, 2010.

[34] Thomas J Fararo and Morris H Sunshine. *A study of a biased friendship net*. Youth Development Center, Syracuse University, 1964.

[35] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, 2009.

[36] Li Fei-Fei. Imagenet: crowdsourcing, benchmarking & other cool things. In *CMU VASC Seminar*, 2010.

[37] Lianli Gao, Jingkuan Song, Feiping Nie, Yan Yan, Nicu Sebe, and Heng Tao Shen. Optimal graph learning with partial tags and multiple features for image and video annotation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4371–4379, 2015.

[38] Yue Gao, Meng Wang, Zheng-Jun Zha, Jialie Shen, Xuelong Li, and Xindong Wu. Visual-textual joint relevance learning for tag-based social image search. *Image Processing, IEEE Transactions on*, 22(1):363–376, 2013.

[39] Peter Gehler and Sebastian Nowozin. On feature combination for multiclass object classification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 221–228. IEEE, 2009.

[40] Scott A Golder and Bernardo A Huberman. Usage patterns of collaborative tagging systems. *Journal of information science*, 32(2):198–208, 2006.

[41] Chen Gong, Dacheng Tao, Stephen J Maybank, Wei Liu, Guoliang Kang, and Jie Yang. Multi-modal curriculum learning for semi-supervised image classification. *IEEE Transactions on Image Processing*, 25(7):3249–3260, 2016.

[42] Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. A multi-view embedding space for modeling internet images, tags, and their semantics. *International journal of computer vision*, 106(2):210–233, 2014.

[43] Yunchao Gong, Svetlana Lazebnik, Albert Gordo, and Florent Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2916–2929, 2013.

[44] Yunchao Gong, Liwei Wang, Micah Hodosh, Julia Hockenmaier, and Svetlana Lazebnik. Improving image-sentence embeddings using large weakly annotated photo collections. In *European Conference on Computer Vision*, pages 529–545. Springer, 2014.

[45] Google. Google books, ngram viewer.

[46] Google. Google cloud platform.

[47] Harry Halpin, Valentin Robu, and Hana Shepherd. The complex dynamics of collaborative tagging. In *Proceedings of the 16th international conference on World Wide Web*, pages 211–220. ACM, 2007.

[48] Ju Han and Kai-Kuang Ma. Fuzzy color histogram and its use in color image retrieval. *IEEE Transactions on image Processing*, 11(8):944–952, 2002.

[49] Derrick Harris. Why video is the next big thing in big data.

[50] Jer Hayes. Multimedia big data: Content analysis and retrieval. In Marcello Trovati, Richard Hill, Ashiq Anjum, Shao Ying Zhu, and Lu Liu, editors, *Big-Data Analytics and Cloud Computing*, pages 37–51. Springer International Publishing, 2015.

[51] Jingrui He and Rick Lawrence. A graph-based framework for multi-task multi-view learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 25–32, 2011.

[52] Mohamed Hefeeda, Tarek ElGamal, Kiana Calagari, and Ahmed Abdelsadek. Cloud-based multimedia content protection system. 2013.

[53] R Hettiarachchi and James F Peters. Multi-manifold lle learning in pattern recognition. *Pattern Recognition*, 48(9):2947–2960, 2015.

[54] Paul Heymann, Georgia Koutrika, and Hector Garcia-Molina. Can social bookmarking improve web search? In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 195–206. ACM, 2008.

[55] Andreas Hotho, Robert Jäschke, Christoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In *European Semantic Web conference*, pages 411–426. Springer, 2006.

[56] Yiqun Hu, Xiangang Cheng, Liang-Tien Chia, Xing Xie, Deepu Rajan, and Ah-Hwee Tan. Coherent phrase model for efficient image near-duplicate retrieval. *IEEE Transactions on Multimedia*, 11(8):1434–1445, 2009.

[57] Yu Hua, Bin Xiao, and Xue Liu. Nest: Locality-aware approximate query service for cloud computing. In *INFOCOM, 2013 Proceedings IEEE*, pages 1303–1311. IEEE, 2013.

[58] Zi Huang, Heng Tao Shen, Jie Shao, Bin Cui, and Xiaofang Zhou. Practical online near-duplicate subsequence detection for continuous video streams. *IEEE Transactions on Multimedia*, 12(5):386–398, 2010.

[59] TaeHyun Hwang and Rui Kuang. A heterogeneous label propagation algorithm for disease gene discovery. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 583–594. SIAM, 2010.

[60] ImageNet. Large scale visual recognition challenge (ilsvrc).

[61] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, STOC '98, pages 604–613, New York, NY, USA, 1998. ACM.

[62] Instagram. Instagram stats.

[63] Chuanjun Ji, Xiangdong Zhou, Lan Lin, and Weidong Yang. Labeling images by integrating sparse multiple distance learning and semantic context modeling. *Computer Vision–ECCV 2012*, pages 688–701, 2012.

[64] Rongrong Ji, Hongxun Yao, Wei Liu, Xiaoshuai Sun, and Qi Tian. Task-dependent visual-codebook compression. *IEEE Transactions on Image Processing*, 21(4):2282–2293, 2012.

[65] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013.

[66] Yangqing Jia, Mathieu Salzmann, and Trevor Darrell. Learning cross-modality similarity for multinomial data. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2407–2414. IEEE, 2011.

[67] Yu-Gang Jiang, Guangnan Ye, Shih-Fu Chang, Daniel Ellis, and Alexander C Loui. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, page 29. ACM, 2011.

[68] Feng Kang, Rong Jin, and Rahul Sukthankar. Correlated label propagation with application to multi-label learning. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1719–1726. IEEE, 2006.

[69] Masayuki Karasuyama and Hiroshi Mamitsuka. Multiple graph label propagation by sparse integration. *IEEE transactions on neural networks and learning systems*, 24(12):1999–2012, 2013.

[70] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137, 2015.

[71] Andrej Karpathy, Armand Joulin, and Fei Fei F Li. Deep fragment embeddings for bidirectional image sentence mapping. In *Advances in neural information processing systems*, pages 1889–1897, 2014.

[72] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.

[73] Yan Ke and Rahul Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–506. IEEE, 2004.

[74] Lyndon S Kennedy, Shih-Fu Chang, and Igor V Kozintsev. To search or to label?: predicting the performance of search-based automatic image classifiers. In *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 249–258. ACM, 2006.

[75] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539*, 2014.

[76] Xiangnan Kong, Bokai Cao, and Philip S Yu. Multi-label classification by mining label and instance correlations from heterogeneous information networks. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 614–622. ACM, 2013.

[77] Xiangnan Kong, Michael K Ng, and Zhi-Hua Zhou. Transductive multilabel learning via label set propagation. *IEEE Transactions on Knowledge and Data Engineering*, 25(3):704–719, 2013.

[78] Christian Körner, Dominik Benz, Andreas Hotho, Markus Strohmaier, and Gerd Stumme. Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. In *Proceedings of the 19th international conference on World wide web*, pages 521–530. ACM, 2010.

[79] Julien Law-To, Olivier Buisson, Valerie Gouet-Brunet, and Nozha Boujemaa. Robust voting algorithm based on labels of behavior for video copy detection. In *Proceedings of the 14th ACM international conference on Multimedia*, pages 835–844. ACM, 2006.

[80] Evan LePage. A long list of instagram statistics and facts (that prove its importance).

[81] Xirong Li, Cees GM Snoek, and Marcel Worring. Learning social tag relevance by neighbor voting. *Multimedia, IEEE Transactions on*, 11(7):1310–1322, 2009.

[82] Xirong Li, Tiberio Uricchio, Lamberto Ballan, Marco Bertini, Cees GM Snoek, and Alberto Del Bimbo. Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. *ACM Computing Surveys (CSUR)*, 49(1):14, 2016.

[83] Rainer Lienhart, Stefan Romberg, and Eva Hörster. Multilayer plsa for multimodal image retrieval. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, page 9. ACM, 2009.

[84] Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang, and Hong-Jiang Zhang. Tag ranking. In *Proceedings of the 18th international conference on World wide web*, pages 351–360. ACM, 2009.

[85] Jiajun Liu, Zi Huang, Hongyun Cai, Heng Tao Shen, Chong Wah Ngo, and Wei Wang. Near-duplicate video retrieval: Current research and future trends. *ACM Computing Surveys (CSUR)*, 45(4):44, 2013.

[86] Li Liu, Yue Lu, and Ching Y Suen. Variable-length signature for near-duplicate image matching. *IEEE Transactions on Image Processing*, 24(4):1282–1296, 2015.

[87] Steffen Lohmann and Paloma Díaz. Representing and visualizing folksonomies as graphs: a reference model. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 729–732. ACM, 2012.

[88] MARINA LOPES. Videos may make up 84 percent of internet traffic by 2018: Cisco.

[89] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[90] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

[91] Benjamin Markines, Ciro Cattuto, Filippo Menczer, Dominik Benz, Andreas Hotho, and Gerd Stumme. Evaluating similarity measures for emergent semantics of social tagging. In *Proceedings of the 18th international conference on World wide web*, pages 641–650. ACM, 2009.

[92] Jesse Mawhinney. 37 visual content marketing statistics you should know in 2016.

[93] Krystian Mikolajczyk, Tinne Tuytelaars, Cordelia Schmid, Andrew Zisserman, Jiri Matas, Frederik Schaffalitzky, Timor Kadir, and Luc Van Gool. A comparison of affine region detectors. *International journal of computer vision*, 65(1-2):43–72, 2005.

[94] Ion Muslea, Steven Minton, and Craig A Knoblock. Active+ semi-supervised learning= robust multi-view learning. In *ICML*, volume 2, pages 435–442, 2002.

[95] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

[96] Zhenhua Nie, Yu Hua, Dan Feng, Qiuyu Li, and Yuanyuan Sun. Efficient storage support for real-time near-duplicate video retrieval. In *Algorithms and Architectures for Parallel Processing*, pages 312–324. Springer, 2014.

[97] Rasmus Pagh and Flemming Friche Rodler. *Algorithms — ESA 2001: 9th Annual European Symposium Århus, Denmark, August 28–31, 2001 Proceedings*, chapter Cuckoo Hashing, pages 121–133. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001.

[98] Jose San Pedro, Stefan Siersdorfer, and Mark Sanderson. Content redundancy in youtube and its application to video tagging. *ACM Transactions on Information Systems (TOIS)*, 29(3):13, 2011.

[99] Jinye Peng, Yi Shen, and Jianping Fan. Cross-modal social image clustering and tag cleansing. *Journal of Visual Communication and Image Representation*, 24(7):895–910, 2013.

[100] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[101] Sujith Ravi and Qiming Diao. Large scale distributed semi-supervised learning using streaming approximation. *arXiv preprint arXiv:1512.01752*, 2015.

[102] Xiang Ren, Yujing Wang, Xiao Yu, Jun Yan, Zheng Chen, and Jiawei Han. Heterogeneous graph-based intent learning with queries, web pages and wikipedia concepts. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 23–32. ACM, 2014.

[103] Valentin Robu, Harry Halpin, and Hana Shepherd. Emergence of consensus and shared vocabularies in collaborative tagging systems. *ACM Transactions on the Web (TWEB)*, 3(4):14, 2009.

[104] Lifeng Shang, Linjun Yang, Fei Wang, Kwok-Ping Chan, and Xian-Sheng Hua. Real-time large scale near-duplicate web video retrieval. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 531–540. ACM, 2010.

[105] Yi Shen and Jianping Fan. Leveraging loosely-tagged images and inter-object correlations for tag recommendation. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 5–14. ACM, 2010.

[106] Stefan Siersdorfer, Jose San Pedro, and Mark Sanderson. Automatic video tagging using content redundancy. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 395–402. ACM, 2009.

[107] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[108] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477. IEEE, 2003.

[109] Jingkuan Song, Yi Yang, Zi Huang, Heng Tao Shen, and Richang Hong. Multiple feature hashing for real-time large scale near-duplicate video retrieval. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 423–432. ACM, 2011.

[110] Jingkuan Song, Yi Yang, Zi Huang, Heng Tao Shen, and Jiebo Luo. Effective multiple feature hashing for large-scale near-duplicate video retrieval. *IEEE Transactions on Multimedia*, 15(8):1997–2008, 2013.

[111] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826, 2016.

[112] Bart Thomee, Mark J Huiskes, Erwin Bakker, and Michael S Lew. Large scale image copy detection evaluation. In *Proceedings of the 1st ACM international conference on Multimedia information retrieval*, pages 59–66. ACM, 2008.

[113] Sergey Tulyakov, Stefan Jaeger, Venu Govindaraju, and David Doermann. Review of classifier combination methods. In *Machine Learning in Document Analysis and Recognition*, pages 361–386. Springer, 2008.

[114] Arash Vahdat and Greg Mori. Handling uncertain tags in visual recognition. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 737–744. IEEE, 2013.

[115] Arash Vahdat, Guang-Tong Zhou, and Greg Mori. Discovering video clusters from visual features and noisy tags. In *Computer Vision–ECCV 2014*, pages 526–539. Springer, 2014.

[116] Bo Wang, Jiayan Jiang, Wei Wang, Zhi-Hua Zhou, and Zhuowen Tu. Unsupervised metric fusion by cross diffusion. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2997–3004. IEEE, 2012.

[117] Bo Wang, Zhuowen Tu, and John K Tsotsos. Dynamic label propagation for semi-supervised multi-class multi-label classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 425–432, 2013.

[118] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1083–1092, 2015.

[119] Xi Wang and Gita Sukthankar. Multi-label relational neighbor classification using social context features. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 464–472. ACM, 2013.

[120] Xiao-Yong Wei, Zhen-Qun Yang, Chong-Wah Ngo, and Wei Zhang. Visual typo correction by collocative optimization: A case study on merchandize images. *IEEE Transactions on Image Processing*, 23(2):527–540, 2014.

[121] Robert Wetzker, Carsten Zimmermann, Christian Bauckhage, and Sahin Albayrak. I tag, you tag: translating tags for advanced user models. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 71–80. ACM, 2010.

[122] Bin Wu, Erheng Zhong, Ben Tan, Andrew Horner, and Qiang Yang. Crowdsourced time-sync video tagging using temporal and personalized topic modeling. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 721–730. ACM, 2014.

[123] Lei Wu, Linjun Yang, Nenghai Yu, and Xian-Sheng Hua. Learning to tag. In *Proceedings of the 18th international conference on World wide web*, pages 361–370. ACM, 2009.

[124] Pengcheng Wu, Steven CH Hoi, Hao Xia, Peilin Zhao, Dayong Wang, and Chunyan Miao. Online multimodal deep similarity learning with application to image retrieval. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 153–162. ACM, 2013.

[125] Xiao Wu, Alexander G Hauptmann, and Chong-Wah Ngo. Practical elimination of near-duplicates from web video search. In *Proceedings of the 15th ACM international conference on Multimedia*, pages 218–227. ACM, 2007.

[126] Xiao Wu, Chong-Wah Ngo, Alexander G Hauptmann, and Hung-Khoon Tan. Real-time near-duplicate elimination for web video search with content and context. *IEEE Transactions on Multimedia*, 11(2):196–207, 2009.

[127] Wensi Xi, Benyu Zhang, Zheng Chen, Yizhou Lu, Shuicheng Yan, Wei-Ying Ma, and Edward Allan Fox. Link fusion: a unified link analysis framework for multi-type interrelated data objects. In *Proceedings of the 13th international conference on World Wide Web*, pages 319–327. ACM, 2004.

[128] Alexander G. Hauptmann Xiao Wu, Chong-Wah Ngo. Cc_web_video: Near-duplicate web video dataset.

[129] Hongtao Xie, Ke Gao, Yongdong Zhang, Sheng Tang, Jintao Li, and Yizhi Liu. Efficient feature detection and effective post-verification for large scale near-duplicate image search. *IEEE transactions on Multimedia*, 13(6):1319–1332, 2011.

[130] Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.

[131] Jiejun Xu, Vignesh Jagadeesh, and BS Manjunath. Multi-label learning with fused multimodal bi-relational graph. *IEEE Transactions on Multimedia*, 16(2):403–412, 2014.

[132] Rong Yan and Milind Naphade. Semi-supervised cross feature learning for semantic concept detection in videos. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 657–663. IEEE, 2005.

[133] Weilong Yang and George Toderici. Discriminative tag learning on youtube videos with latent sub-tags. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3217–3224. IEEE, 2011.

[134] Yang Yang, Yi Yang, and Heng Tao Shen. Effective transfer tagging from image to video. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 9(2):14, 2013.

[135] Ting Yao, Tao Mei, Chong-Wah Ngo, and Shipeng Li. Annotation for free: Video tagging by mining user search behavior. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 977–986. ACM, 2013.

[136] YouTube. Statistics.

[137] YouTube. Statistics.

[138] Jun Yu, Yong Rui, and Dacheng Tao. Click prediction for web image reranking using multimodal sparse coding. *IEEE Transactions on Image Processing*, 23(5):2019–2032, 2014.

[139] Dongqing Zhang and Wu-Jun Li. Large-scale supervised multimodal hashing with semantic correlation maximization. In *AAAI*, pages 2177–2183, 2014.

[140] Hong Jiang Zhang, Jianhua Wu, Di Zhong, and Stephen W Smoliar. An integrated system for content-based video retrieval and browsing. *Pattern recognition*, 30(4):643–658, 1997.

[141] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837, 2014.

[142] Qi Zhang, Sally A Goldman, Wei Yu, and Jason E Fritts. Content-based image retrieval using multiple-instance learning. In *ICML*, volume 2, pages 682–689. Citeseer, 2002.

[143] Rui Zhang, Lei Zhang, Xin-Jing Wang, and Ling Guan. Multi-feature plsa for combining visual features in image annotation. In *Proceedings of the 19th ACM international conference on Multimedia*, pages 1513–1516. ACM, 2011.

[144] Shiliang Zhang, Qi Tian, Gang Hua, Qingming Huang, and Wen Gao. Generating descriptive visual words and visual phrases for large-scale image applications. *IEEE Transactions on Image Processing*, 20(9):2664–2677, 2011.

[145] Wan-Lei Zhao and Chong-Wah Ngo. Scale-rotation invariant pattern entropy for keypoint-based near-duplicate detection. *IEEE Transactions on Image Processing*, 18(2):412–423, 2009.

[146] Wan-Lei Zhao, Chong-Wah Ngo, Hung-Khoon Tan, and Xiao Wu. Near-duplicate keyframe identification with interest point matching and pattern learning. *IEEE Transactions on Multimedia*, 9(5):1037–1048, 2007.

[147] Xiangmin Zhou and Lei Chen. Monitoring near duplicates over video streams. In *Proceedings of the 18th ACM international conference on multimedia*, pages 521–530. ACM, 2010.

[148] Xiangmin Zhou, Lei Chen, and Xiaofang Zhou. Structure tensor series-based large scale near-duplicate video retrieval. *IEEE Transactions on multimedia*, 14(4):1220–1233, 2012.

[149] Xiangmin Zhou, Xiaofang Zhou, Lei Chen, Athman Bouguettaya, Nong Xiao, and John A Taylor. An efficient near-duplicate video shot detection method using shot-based interest points. *IEEE Transactions on Multimedia*, 11(5):879–891, 2009.

[150] Shiai Zhu, Chong-Wah Ngo, and Yu-Gang Jiang. Sampling and ontologically pooling web images for visual concept learning. *Multimedia, IEEE Transactions on*, 14(4):1068–1078, 2012.

[151] Xiaofeng Zhu, Zi Huang, Heng Tao Shen, and Xin Zhao. Linear cross-modal hashing for efficient multimedia search. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 143–152. ACM, 2013.

[152] Xiaojin Zhu and Zoubin Ghahramani. Learning from labeled and unlabeled data with label propagation. 2002.

[153] Xiaojin Zhu, Zoubin Ghahramani, and John D Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003.

[154] Yue Ting Zhuang, Yan Fei Wang, Fei Wu, Yin Zhang, and Wei Ming Lu. Supervised coupled dictionary learning with group structures for multi-modal retrieval. In *Twenty-Seventh AAAI Conference on Artificial Intelligence*, 2013.

[155] Fuhao Zou, Yunpeng Chen, Jingkuan Song, Ke Zhou, Yang Yang, and Nicu Sebe. Compact image fingerprint via multiple kernel hashing. *IEEE Transactions on Multimedia*, 17(7):1006–1018, 2015.