A Review of Video Retrieval Based on Image and Video Semantic Understanding

Miki Haseyama (member)†, Takahiro Ogawa (member)†, Nobuyuki Yagi (member)††

Abstract Research trends in new video retrieval based on image and video semantic understanding are presented in this paper. First, recent studies related to image and video semantic analysis are introduced to understand leading-edge multimedia retrieval technologies. Several works related to visualization interfaces for multimedia retrieval are also presented. Finally, trends in state-of-the-art studies and the future outlook are described.

Key words: video retrieval, image and video semantic understanding, visualization, navigation.

1. Introduction

According to the IDC White Paper published in March 2008, the digital universe has reached 281 exabytes and has exceeded the amount of available worldwide storage for the first time\(^1\). It was also reported that not all information created and transmitted gets stored — almost half of the digital data would not have a permanent home in 2011. Furthermore, according to the IDC report published in May 2010, the digital universe is expected to be 44-times larger in 2020 than in 2009\(^2\). It was also reported that much of the data included in the digital universe is unstructured data such as images, music and videos, i.e., multimedia data. Thus, the amount of data on the Web is huge, and we are faced with a new problem: how do we find information that we need when we need it? This problem is closely related to new tools for retrieval and discovery.

Currently, users generally use the query-response model to access their desired information on the Web. In the query-response model, the user submits a query through an interface to obtain the desired information. According to the query, this model narrows the huge amount of information down to an amount that the user can check. In this case, the multimedia contents being searched have been annotated with metadata in advance, and the search engine presents the user with the contents whose metadata match the query. Conventionally, metadata include the time and location, which are attached to the content at the time of acquisition, and manually attached keywords.

In recent years, progress in research of image and video analysis, particularly image and video semantic understanding, has led to the development of technologies that automatically extract metadata\(^3\)\(^4\). Machine learning has contributed tremendously to the development of these technologies, especially for deriving semantic level indices, and it is accelerating the progress in content-based image and video retrieval research\(^5\). Therefore, trends in these research areas are important for understanding recent trends in multimedia retrieval.

Although the accuracy of image and video semantic understanding is increasing, it is still difficult to perfectly solve the problem of the “Semantic Gap”\(^6\). Specifically, although recent machine learning-based methods can annotate multimedia contents with keywords, they require sufficient amounts of training data, without which highly accurate annotation becomes difficult. In general, it is difficult to provide sufficient training data in all cases, and the accuracy often becomes insufficient. In such cases, since it is difficult to accurately attach the metadata to contents, it also becomes difficult to perform accurate retrieval by calculating inter-content distances or similarities based on metadata.

In addition to the above problems, a new problem has recently emerged. In some cases, we are unable to provide a query that accurately represents the desired content (e.g., keyword, image, etc), and in such cases, finding the desired content through the query-response model becomes difficult\(^7\). This problem was also pointed out in the IDC report in 2010\(^8\). It was also stated in that report that we will have to consider how we find the information we need when we need...
it, and thus new retrieval and discovery tools must be
developed. Recently, in order to solve this problem,
several studies on “human-centered computing”^5,6^ have
been carried out. In human-centered computing, the
main idea is to satisfy users and allow users to make
queries in their own terminology. User studies have pro-
vided direct insights into interactions between humans
and computers. By human-centered, we mean systems
that consider the behavior and needs of the human user.
The foundational areas of multimedia information re-
trieval were often in computing-centric fields. Specific
examples in this research area include a visualization
interface that connects the user and the computer. This
is known as the method that provides the most effec-
tive solution and is becoming an essential component
for realizing more efficient multimedia retrieval.

In this paper, we present research trends in new video
retrieval based on image and video semantic un-
derstanding. Figure 1 shows an overview of the recent
research trends for video retrieval based on image and
video semantic understanding. First, we provide an
explanation of existing research related to image and
video semantic understanding, which is important for
understanding multimedia retrieval. Furthermore, we
present an introduction of research trends related to
visualization interfaces for multimedia retrieval, which
have attracted attention in recent years. The first half
of this paper provides details of image and video se-
matic understanding. Many image and video repre-
sentation methods, that is, methods for calculating fea-
tures that can accurately represent contents, have been
proposed. Therefore, achievements of methods that
have been reported in recent years are described. Fur-
thermore, since collectively investigating classification
methods based on extracted features are necessary for
image and video semantic understanding, we provide a
broad explanation ranging from fundamental methods
to the latest machine learning-based methods. In the
second half of this paper, details of visualization inter-
faces for multimedia content retrieval are provided. In
addition to achievements of existing studies, recent re-
search trends and the future outlook are described in
this paper.

2. Image and Video Semantic Under-
standing

As described above, image and video retrieval gen-
erally becomes feasible by annotating contents with
metadata. Researchers foresaw the arrival of the age
of large-scale image and video information retrieval^9^,
and research of image and video understanding began.
Many of these methods focused on low-level features
and enabled retrieval of images in which color or tex-
ture features are similar to those of the provided query.
However, in order to realize retrieval of content that the
user truly desires, high-level features are necessary, and
attaching metadata that represent the meaning of the
contents is essential. Although in the past, it was con-
sidered that realization of such technologies is difficult,
corpus-based multimedia analysis methods have been
proposed in recent years, and image and video semantic
understanding technologies have advanced significantly
toward practical use. Examples of leading corpus-based
analysis technologies include speech recognition, natu-
ral language processing and character recognition. In
the field of image and video analysis, by adopting the
above approach, methods that can cope flexibly with
the diversity and ambiguity of multimedia contents
have been realized. The contribution of machine learn-
ing is significant. Many features that are effective for
image and video semantic understanding have also been
proposed. The recent leading methods tend to be ex-
tended from those proposed in the field of language pro-
cessing. These methods are also expected to be highly
compatible with statistical machine learning methods.

In this section, methods for representing image and
video contents, in other words, recent research trends
related to visual features are first explained in Sec. 2.1.
Sec. 2.2 provides a broad explanation of statistical mod-
els and classifiers utilized for image and video semantic
understanding, ranging from fundamental methods to
the latest methods.

2.1 Image and Video Representation Meth-
ods

Generally, visual features used to represent image and
video contents can be broadly divided into two main
classes: global features and local features. Global fea-
tures are calculated from the entire image and include
color histograms^9^, HLAC^10^, Gist^11^, etc. These fea-
tures have been studied since the 1990s and used in
content-based image retrieval (CBIR). They realized
the retrieval of similar images by searching for simi-
lar features in a database. These methods are still
used as standard retrieval methods in image and video
databases. However, since they utilize low-level fea-
tures that do not accurately represent the contents, re-
trieval, which focuses on the objects included in image
and video contents, becomes difficult.

From the late 1990s until the mid 2000s, the use of local features led to a breakthrough that solved the above problem\(^{12}\). The most representative use of local feature is SIFT (Scale Invariant Feature Transform) proposed by Lowe\(^{13-14}\). This method was proposed for object detection in the presence of an occlusion and provides a new framework for the extraction of feature points and their descriptions. Other methods derived from SIFT, such as PCA-SIFT\(^{15}\), which performs dimensionality reduction of gradient information via PCA, BSIFT\(^{16}\), which reduces the effect of background information, and SURF (Speeded Up Robust Feature)\(^{17}\) have recently been proposed. HOG (Histograms of Oriented Gradients)\(^{18}\), a feature descriptor for a localized region, has also been proposed. This is widely used in general object recognition, e.g., recognition of humans or vehicles. In the past few years, various methods for performing hierarchical matching based on the above features have also been proposed\(^{19-20}\).

Although the above local features enable accurate and robust description of image and video contents, comparing feature points from many contents becomes difficult when a large number of local features is extracted from a single content. Therefore, a method that retrieves corresponding feature points by quantizing the local feature vector extracted from a single content and replacing it with a codeword that is a representative vector (visual word) has been proposed\(^{21}\). The bag-of-features (BoF) representation\(^{22}\), which allowed application of visual words from specific object recognition to general object recognition, was also proposed and it has become one of the most widely used features in recent years. BoF representation is similar to bag-of-words (BoW) representation, which describes a text document as a vector by ignoring word order and only focusing on the frequency of word occurrence, and it is used in the fields of language processing and information retrieval. BoF representation ignores the location of each feature point and applies the bag-of-words method to visual words. Recently, many extensions of the BoF method have been proposed. Specifically, a method that provides a codebook of visual words based on online clustering and mean-shift\(^{23}\), a method based on probabilistic clustering through GMM (Generalized Method of Moments) and the EM (Expectation-Maximization) algorithms\(^{24}\), and a method utilizing color and spatiotemporal information\(^{25}\) have been proposed. In the most recent research, a method that utilizes locality-constrained linear coding\(^{26}\) has also been used as an extension of BoF. In addition, bag-of-video-words, a method that applies BoF to video contents, has also been proposed\(^{27}\). An overview of the trends in image and video representation methods is presented in “Visual Features” of Fig. 1.

### 2.2 Statistical Models and Classifiers

Generally, after the extraction of features representing image and video content (explained in the previous subsection), the content is classified by some statistical models or classifiers. In the past, probabilistic models for performing document classification, such as probabilistic Latent Semantic Analysis (pLSA)\(^{28}\) and Latent Dirichlet Allocation\(^{29}\), have been proposed. These methods are also applied to general object recognition in image and video contents. Similar to the application of BoW representation to BoF representation, these methods are notable examples of adopting methods from the field of language processing. Latent Semantic Analysis (LSA)\(^{30}\) realizes topic extraction by performing singular value decomposition over a collection of feature vectors, and pLSA\(^{28}\), its probabilistic extension, enables the extraction of latent topics. In recent years, LDA\(^{29}\) has been proposed as an expansion of pLSA. Methods that model relationships between heterogeneous features as latent variables (corresponding to latent topics), such as probabilistic canonical correlation analysis (PCCA)\(^{31}\), have also been proposed.

In contrast to the above methods, various classification methods that adopt different approaches based on multivariate analysis of the obtained set of features have been proposed. The subspace method\(^{32}\), the most fundamental method, enables classification by calculating the similarity between the input features and the subspace spanned by the features of each class. In addition, when the input is provided as multiple features, the mutual subspace method\(^{33}\) enables classification by comparing the subspace of the input and that of each class. Furthermore, the constrained mutual subspace method\(^{34}\) and the whitened mutual subspace method\(^{35}\) were proposed to solve the problem of accuracy degradation due to the overlap between the subspaces of different classes. By using nonlinear subspaces based on kernel principal component analysis\(^{36}\), those methods were respectively extended to the kernel subspace method\(^{37}\), the kernel mutual subspace method\(^{38}\), the kernel constrained subspace method\(^{39}\) and kernel orthogonal mutual subspace method\(^{40}\).
SVM$^{41}$ has recently become the most well-known machine learning classifier, and it is often used in classification of video content. Support Vector Data Description (SVDD)$^{42}$ and One-Class SVM (OCSVM)$^{43}$ are also known as classifiers for one-class problems, where one class problem tries to distinguish one class of objects from all other possible objects. These classifiers have been actively studied, and various extended methods exist. Specifically, the use of kernel methods with classifiers is important for improving classification accuracy, and it is necessary to select the best kernel that is optimal to the subject of application$^{44}$. Also, various methods that aim to improve accuracy by combining multiple kernels have been proposed$^{45,46}$. In contrast to the above-mentioned methods that construct strong classifiers, methods that realize highly accurate classification by combining weak classifiers trained from large amounts of training data have recently been proposed, with the most notable method being Ada-boost$^{47}$. This method is used in human and face classification, and the VJ (Viola & Jones) method has become one of the most frequently used methods$^{48}$. An overview of the research trends in statistical models and classifiers is presented in “Model and Classifiers” of Fig. 1.

3. Visualization Interfaces for Image and Video Retrieval

As shown in the previous section, recent developments in image and video semantic understanding have contributed significantly to improvement in the accuracy of image and video retrieval technology. It should be noted that existing systems that perform retrieval are based on the Query-Response model, and there is a large number of systems displaying search results based on the similarity between query content and contents in the database. However, it becomes difficult for those systems to provide useful information to users due to the semantic gap or in cases in which users cannot provide accurate queries. Note that the semantic gap generally means the difference between visual features extracted from image and video data and high-level information of humans. In recent years, general studies related to human-centered computing$^{49,50}$ have been performed to address this problem. Visualization systems (also known as navigation systems)$^{51}$ have been developed and are expected to become an alternative to the Query-Response model.

3.1 Existing Research Trends

Many visualization methods for image and video retrieval have been proposed. As shown in “Visualization Interface” of Fig. 1, conventional methods for visualizing large amounts of retrieved contents are broadly divided into three main classes: dimensionality reduced mappings, clustering-based visualizations and graph-based representations$^{52,53}$. Methods based on dimensionality reduced mappings calculate the similarities between contents in a high-dimensional feature space and represent them in lower-dimensional spaces, i.e., visible dimensional (2D or 3D) spaces. Then users can easily understand the similarities between the visualized contents. Since these methods perform a mapping from a high-dimensional feature space to a low-dimensional feature space, conventional dimensionality reduction methods from the field of multivariate analysis are often adopted. A browsing method$^{54}$ based on dimensionality reduction using PCA, an MDS (Multi-Dimensional Scaling)$^{55}$-based browsing method$^{56}$, and also an SOM (Self-Organized Maps)$^{57}$-based method$^{58}$ in which SOM enables non-linear mappings have been proposed. Furthermore, in the past ten years, several methods based on ongoing research related to Manifold Learning, such as ISOMAP (isometric mapping)$^{59}$, SNE (stochastic neighbor embedding)$^{60}$ and LLE (locally linear embedding)$^{61}$, have been proposed.

In clustering-based visualization, in order to enable users to instantly grasp the entire contents being searched, the amount of contents is reduced by grouping similar contents. In this case, content similarity can be calculated on the basis of data such as features extracted from the image contents$^{62,63}$, metadata$^{64}$ and the content acquisition time$^{65}$. Then users can retrieve the desired contents by narrowing down from the representative contents of each cluster.

On the other hand, graph-based representations enable the visualization of content collection by constructing a graph. In these approaches, the contents are regarded as nodes, and two nodes with a high similarity are connected by edges. Furthermore, in these approaches, similarities are calculated on the basis of content’s features$^{66}$ and shared-keyword annotation$^{67}$. In addition to the above three broad classes, various visualization methods have been proposed, and methods that share different classes also exist$^{68,69}$. Many good surveys of image and video visualization and navigation have been published$^{53,70,71}$.72}
3.2 Recent Research Trends and Future Outlook

As shown in the previous subsection, the use of visualization can overcome the limitations of conventional retrieval systems. By developing mutually, the fields of image and video semantic understanding and visualization have enabled the provision of higher-level multimedia content retrieval to users. Therefore, as shown in “Recent Trends” of Fig. 1, research areas that realize acceleration for each field have progressed rapidly in recent years. Specifically, an extremely large number of methods that perform image and video semantic understanding and visualization follow the multimodal approach\(^{(79)}\). Video content includes visual and audio data (and sometimes includes closed captions), and it has been shown that their collaborative use can lead to improvement of retrieval accuracy. In practice, there are various methods that realize the improvement of retrieval accuracy by introducing a multimodal approach\(^{(74,77,78)}\). In the field of image and video semantic understanding, there have also been attempts to apply multimodal processing to statistical models or classifiers\(^{(77)}\). When targeting content on the Web, new approaches that use text data from the Web page including that content and inter-page link relationships have also been proposed\(^{(79)}\).

In the field of image and video retrieval, including visualization, some trials to improve retrieval accuracy through interactions with the user have also been carried out. The leading method is RF (Relevance Feedback), which utilizes feedback from the user to interactively bring the search results closer to the desired content\(^{(79)}\). In practice, the RF method has been actively introduced to the field of video retrieval and is used in several methods shown above\(^{(75,76,78)}\). Furthermore, through a similar way of thinking, Associative Search\(^{(80)}\) used with text data has been introduced to video retrieval\(^{(81)}\). Even if users cannot provide accurate queries, these methods repetitively select a new query from the retrieval results, gradually increasing retrieval accuracy and finally leading to the desired information. A good survey on Interactive Search is provided in a reference\(^{(82)}\) since a number of approaches are introduced.

As described above, the fields of image and video semantic understanding and visualization are very closely related to each other in multimedia retrieval. However, there has been a tendency for them to develop as independent fields. The birth of a new research area that crosses these fields is expected in the future.

4. Conclusion

In this paper, research trends related to new video retrieval based on image and video semantic understanding have been presented. Recent trends related to visualization interfaces for new multimedia retrieval that have been researched in recent years have also been introduced. By applying multimedia technology, which is realized by fusing various research fields, new services will be created. In the future, it is expected that a diverse range of services will be developed.

References


Miki Haseyama received her B.S., M.S. and Ph.D. degrees in Electronics from Hokkaido University, Japan in 1986, 1988 and 1993, respectively. She joined the Graduate School of Information Science and Technology, Hokkaido University as an associate professor in 1994. She was a visiting associate professor of Washington University, USA from 2005 to 2006. She is currently a professor in the Graduate School of Information Science and Technology, Hokkaido University. Her research interests include image and video processing and its development into semantic analysis. She is a member of the IEEE, IEICE, Institute of Image Information and Television Engineers (ITF) and Acoustical Society of Japan (ASJ).

Takahiro Ogawa received his B.S., M.S. and Ph.D. degrees in Electronics and Information Engineering from Hokkaido University, Japan in 2003, 2005 and 2007, respectively. He is currently an assistant professor in the Graduate School of Information Science and Technology, Hokkaido University. His research interests are digital image processing and its applications. He is a member of the IEEE, EURASIP-IEICE, and Institute of Information and Television Engineers (ITF).

Nobuyuki Yagi is a professor of Tokyo City University. He received the B.E., M.E., and Ph.D. degrees in electrical engineering from Kyoto University, Japan, in 1978, 1980, and 1992. He joined NHK in 1980, and worked at the Kofu Broadcasting Station, the Science and Technology Research Laboratories, the Engineering Administration Department and Programming Department of NHK. He moved to Tokyo City University in 2012. He was also an affiliate professor of the Tokyo Institute of Technology from 2005 to 2008. His research interests include image and video signal processing, multimedia processing, content production technology, computer architecture, and digital broadcasting. He had contributed to standardization activities at ITU, SMPTE, EBU, ABU, and ARIB. He is a member of IEEE, IEICE, IPSJ, and ITEJ.
Overview of recent research trends for video retrieval based on image and video semantic understanding.