# Instance Search Technology for Finding Specific Objects in Movies

*Masaya Murata, Hidehisa Nagano, Ryo Mukai, Kaoru Hiramatsu, and Kunio Kashino*

## Abstract

Successful retrieval of multimedia such as images or videos often involves utilizing their textual metadata. However, adding such metadata information to all multimedia of interest is far beyond our capability. We are therefore pursuing an effective and efficient content-based search methodology based solely on the multimedia content itself. Our research group has been actively advancing multimedia identification technologies for more than 15 years, and our latest search method makes it possible to search for and find specific objects in many kinds of movies. A specific object, person, or place, is called an *instance*, and the retrieval of an object is called an *instance search*. In this article, we overview our instance search methodology.

*Keywords: video retrieval, instance search, robust media search*

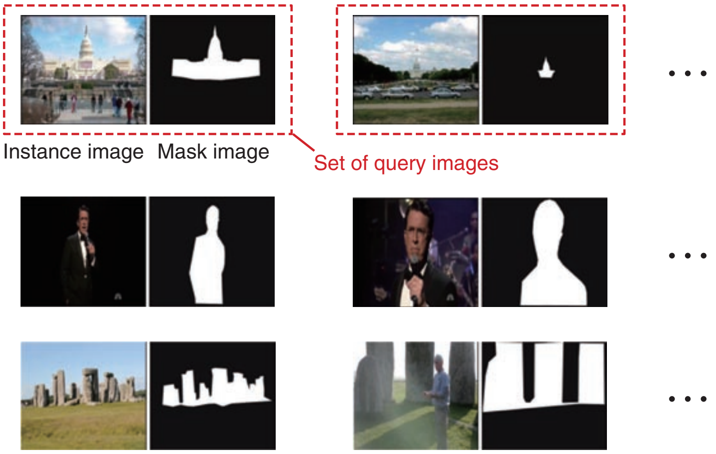
## 1. Introduction

### 1.1 Instance search task

Our research group has been conducting research on multimedia search technology for music, images, and video since the 1990s. Initially, we focused on identifying the same multimedia content, that is, on finding the same signal as the input signal. We believed that such technology would contribute to establishing the fundamentals for future media processing technology, in the same way that term (word) identification techniques have done in natural language processing. Our identification technology, called robust media search (RMS), enables the exact specification of multimedia content with high speed and high accuracy.

The research objective of our RMS team was to find the same signal as the input signal. Our BAM (binary area matching) and DAL (divide and locate) methods made this possible even if signals were corrupted with significant noise and distortion. This technical feature is known as robustness. With this innovative high-speed processing technique, RMS is now driving many services such as music search on mobile phones, music rights management for broadcasting and Internet distribution, World Wide Web (WWW) content identification, and content audience surveys. The upcoming instance search technology has enormous potential to advance RMS as a way to solve challenging real-world problems.

After completing the RMS product development, we focused next on not only finding the same signal, but also finding the same object in multimedia content. For example, the aforementioned instance search technology is aimed at searching for the same object such as a person, logo, or place, as the input instance in movies that have a different background and appearance (**Fig. 1**). It is expected that such search technology will become a powerful tool for organizing large-scale image and video data and for retrieving various kinds of information using real-world image queries. It is further expected that the instance search will realize automatic annotation of a specific person to any scene in video archives and that the information retrieval will be achieved by simply taking photos of unknown objects in, for example, a town or city. The potential of instance search is not limited to these kinds of applications. For example, it may contribute to the establishment of a multimedia dictionary. Such a dictionary would enable more



U.S. Capitol building (object instance, top panels), Stephen Colbert (person instance, center panels), Stonehenge (place instance, bottom panels). These examples are actual query images evaluated at an international workshop called TRECVID (see body of article for more details). At TRECVID, mask images showing the instance regions within the query images are also provided; the white regions indicate where the instance appears within the query images.

Fig. 1. Example sets of query images.

detailed semantic analyses of multimedia content.

### 1.2 Technical aspects of instance search

To find a specific object with a different appearance on a different background, we use local feature data extracted from a query image and a video keyframe. RMS takes a similar approach such as using local features of signals, but it is based on the fact that between the same signals, similar local features tend to occur at the same position in the spatiotemporal space. However, in cases where an object's appearance and background might differ, such consistency is not generally expected, which makes the instance search task very challenging. Our method tackles this problem by precisely matching local features extracted from a query image and a video keyframe in order to differentiate the same object from similar ones. We also consider the discriminative power of local features and assign larger weights to features that contribute to successful instance searching. In short, accurate matching of a highly discriminative local feature is the key to carrying out the instance search task.

## 2. Keypoint based search methodology

Our proposed instance search methodology is illustrated in **Fig. 2** [1]. We first detect characteristic keypoints from query images and video keyframes. These keypoints are respectively called query keypoints and video keypoints. We then describe each keypoint using a high-dimensional vector. We currently employ the Harris-Laplace detector for the keypoint extraction, and we use the scale-invariant feature transform (SIFT) and color SIFT for the feature vector description. Then we find the video keypoint closest to each query keypoint with the criterion that the cosine similarity value between the two vectors is greater than 0.9. The keypoint matching results enable us to count how many times a query keypoint occurs in a video. This number is called a *keypoint frequency*, and by taking all of the frequencies into account, we can determine whether the query instance appears in the video.

However, as mentioned in the previous section, there are two kinds of query keypoints: those with high discriminative power and those with low discriminative power, and we also consider these properties for ranking videos stored in the database.

Such kinds of discriminative power were originally proposed as inverse document frequency (IDF) in the text retrieval field, and the state-of-the-art ranking method called BM25 theoretically validates the use of the IDF weight for ranking documents of interest. The IDF weight quantifies the discriminative power of a keyword term in a document; the value is smaller when the term tends to occur in many kinds of documents. However, for the video retrieval task, we

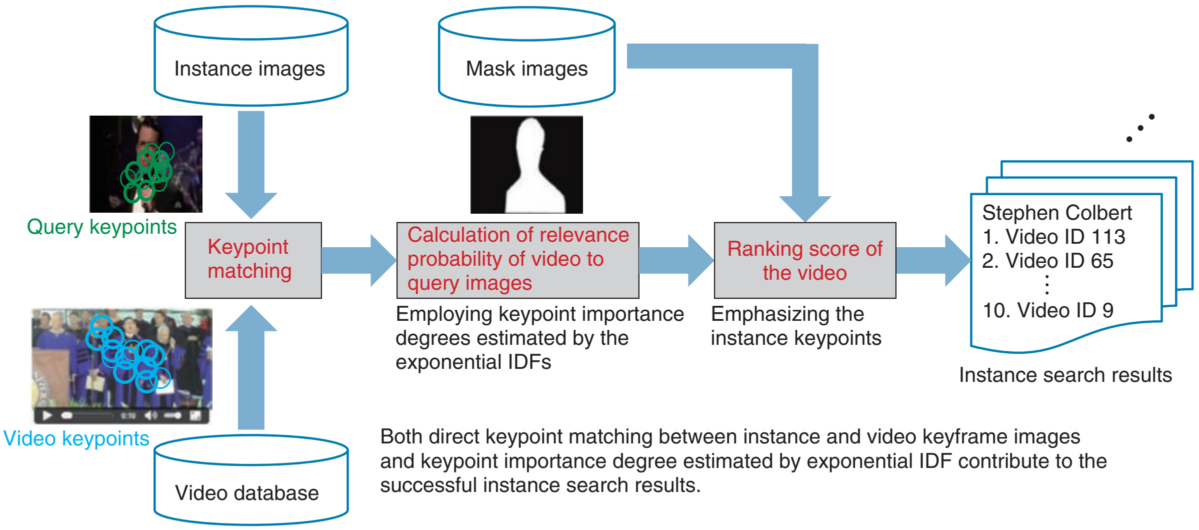

Instance images
Mask images
Query keypoints
Keypoint matching
Calculation of relevance probability of video to query images
Ranking score of the video
Employing keypoint importance degrees estimated by the exponential IDFs
Emphasizing the instance keypoints
Stephen Colbert
1. Video ID 113
2. Video ID 65
⋮
10. Video ID 9
Instance search results
Video keypoints
Video database
Both direct keypoint matching between instance and video keyframe images and keypoint importance degree estimated by exponential IDF contribute to the successful instance search results.

Fig. 2. Overview of instance search method.

found that the conventional IDF could not sufficiently lower the weights of lowly discriminative keypoints. The main reason for this is the inconsistency between the assumptions behind the IDF formulation and the actual properties/features of the images used for instance search.

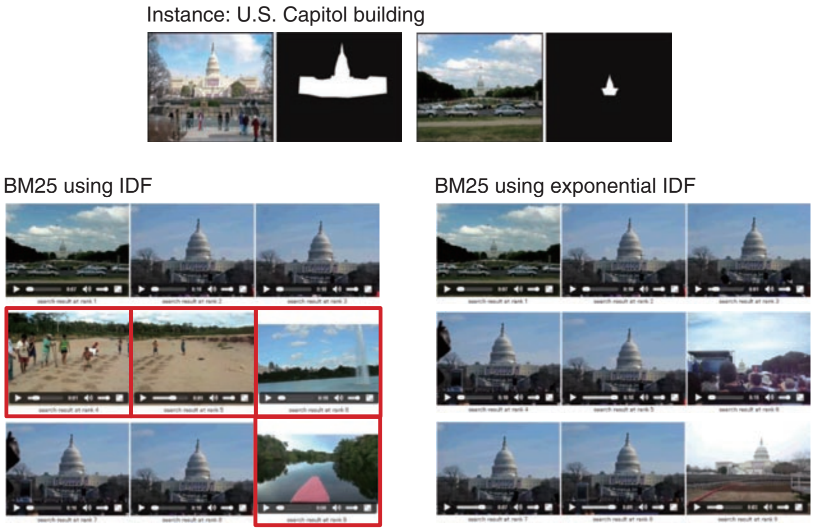
Therefore, we proposed a new keypoint weight that is suitable for the instance search task [2]. We call it exponential IDF, and it is designed to become sufficiently small when the keypoint shows a tendency to be a lowly discriminative feature. The BM25 using exponential IDF realizes accurate matching of highly discriminative local features and results in a high accuracy instance search. More accurate retrieval can generally be achieved by emphasizing keypoints extracted from the specified instance region in the mask image (as in Fig. 1), if the mask image is available. The videos in the database are ranked in decreasing order according to the BM25 scores with exponential IDF and shown to the search user as the instance search result.

The effect of exponential IDF is shown in **Fig. 3**. The uppermost panels correspond to the query and mask images, and the lower left and right search results are those obtained by using the standard IDF and exponential IDF. As seen in the figure, the standard IDF resulted in misdetection caused by the sky, sand, and trees in the query images, but the proposed method successfully suppressed such contributions from the lowly discriminative keypoints.

## 3. System and evaluation

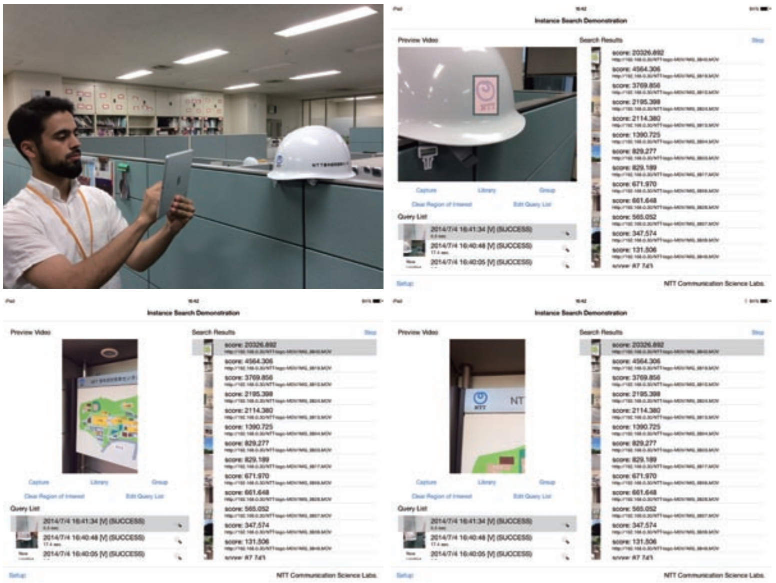
An example of how our instance search system works is shown in **Fig. 4**. The user (top left) takes a photo of an object for a search using a tablet device. Here, the instance is the NTT logo on a white helmet. The user also specifies the region of interest within the query image (top right). Touching the search button initiates the instance search, and the video ranking results are provided on the right side of the tablet device (bottom left). A video showing the NTT logo can be previewed by tapping the thumbnail image on the video result rankings. If the video contains textual metadata information, the user can be directed to an outside information source such as the WWW and achieve the information retrieval by simply taking a photo.

The instance search task is attracting attention from video retrieval researchers, and it is indeed one of the important tasks at the TRECVID (TREC (Text Retrieval Conference) Video Retrieval Evaluation) workshop organized by the U.S. National Institute of Standards and Technology. Instance examples shown in Figs. 1 and 3 are the actual queries used in the previous TRECVID instance search tasks. At the TRECVID workshop held in 2013, our approach recorded the highest-level instance search accuracy among the 23 teams participating from all over the world [3]. This success is mainly because of our keypoint weighting technique described in section 2.



IDF often fails to estimate appropriate importance degrees of query keypoints, resulting in high rankings for incorrect videos (outlined in red).

Fig. 3. Comparison of instance search results.



score: 20326.892
score: 4564.306
score: 3769.856
score: 2195.398
score: 2114.380
score: 1390.725
score: 829.277
score: 829.189
score: 671.970
score: 661.648
score: 565.052
score: 347.574
score: 131.506

2014/7/4 16:41:34 [V] (SUCCESS)
2014/7/4 16:40:48 [V] (SUCCESS)
2014/7/4 16:40:05 [V] (SUCCESS)

(Top left) Taking a photo of the NTT logo on a white helmet. (Top right) Specifying the instance region within the query image; the video search results are shown on the right side of the tablet screen. (Bottom left) Viewing the top ranked video search result by tapping it. (Bottom right) The NTT logo appears in the video shown on the left part of the screen.

Fig. 4. Developing a demo system.

## 4. Future directions

We introduced our approach for an instance search task and focused in particular on the newly devised keypoint weighting method. We will continue to investigate the instance search task from various viewpoints and dedicate ourselves to establishing robust media search technology.

**Masaya Murata**

Research Scientist, Media Recognition Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received his B.S. and M.S. in physics from Waseda University, Tokyo, in 2005 and 2007, respectively. He joined NTT in 2007. In 2008, he received a best paper award at the 9th international conference on Web Information Systems Engineering (WISE2008). Since October 2014, he has also been a Ph.D. student at Graduate School of Information Science and Technology, the University of Tokyo. His area of research includes information retrieval, state filtering, and multimedia search applications.

**Hidehisa Nagano**

Senior Research Scientist, Media Recognition Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received his B.Eng. and M.Eng. in information and computer sciences in 1994 and 1996, respectively, and his Ph.D. in information science and technology in 2005, all from Osaka University, Japan. He joined NTT in 1996. From 2011 to 2012, he was a visiting researcher at the Centre for Digital Music, Queen Mary University of London, UK. He has been working on audio and video analysis, search, retrieval, and recognition algorithms and their implementation. He is a senior member of the Institute of Electrical and Electronics Engineers (IEEE) and a member of the Institute of Electronics, Information and Communication Engineers (IEICE) and the Information Processing Society of Japan (IPSJ).

**Ryo Mukai**

Senior Research Scientist, Media Recognition Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received his B.S. and M.S. in information science from the University of Tokyo in 1990 and 1992, respectively. He joined NTT in 1992. From 1992 to 2000, he was engaged in the research and development of processor architecture for network service systems and distributed network systems. Since 2000, he has been with NTT Communication Science Laboratories, where he is conducting research on media search technology. He is a senior member of IEEE and a member of the Association for Computing Machinery, the Acoustical Society of Japan, IEICE, and IPSJ.

**Kaoru Hiramatsu**

Senior Research Scientist, Supervisor, Media Recognition Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received his B.S. in electrical engineering and his M.S. in computer science from Keio University, Tokyo, in 1994 and 1996, respectively, and his Ph.D. in informatics from Kyoto University in 2002. Since joining NTT Communication Science Laboratories in 1996, he has been working on the Semantic Web, sensor networks, and media search technology. From April 2003 to March 2004, he was a visiting research scientist at the Maryland Information and Network Dynamics Laboratory, University of Maryland, USA. He is a member of IPSJ and the Japanese Society of Artificial Intelligence.

**Kunio Kashino**

Distinguished Researcher, Group Leader of Media Recognition Research Group, Media Information Laboratory, NTT Communication Science Laboratories.

He received his Ph.D. from the University of Tokyo in 1995. Since joining NTT in 1995, he has been working on audio and video analysis, synthesis, search, and recognition algorithms and their implementation. He has received several awards including the Maejima Award in 2010, the Young Scientists' Prize for Commendation for Science and Technology from the Minister of Education, Culture, Sports, Science and Technology in 2007, and the IEEE Transactions on Multimedia Paper Award in 2004. He was a visiting scholar at the University of Cambridge, UK, in 2002 and has been a visiting professor at the National Institute of Informatics, Tokyo, since 2008. He is a senior member of IEEE.