AN INTERFACE FOR IMAGE RETRIEVAL AND ITS EXTENSION TO VIDEO RETRIEVAL

Xây dựng một giao diện cho cho tìm kiếm ảnh và mở rộng cho tìm kiếm video

Lê Thị Lan, Alain Boucher, Monique Thonnat

Abstract

Semantic video retrieval is still an open problem. While many works exist in analyzing the video contents, few ones present the retrieval results to the users and interact with him/her. In this article, firstly, we propose a 2D graphic interface adapted to the problem of image retrieval that enables a bidirectional communication: from the system towards the user to visualize the current research results and from the user towards the system so that the user can provide some relevance feedback information to refine his/her query. In this interface, the visualization shows the image query in the middle of the screen and the result images in a 2D plan with distances showing the similarity measures between images and the query. We propose also a method of relevance feedback in form of validation, in this interface, for image retrieval. This approach has been implemented and tested with different image databases. Secondly, we analyze the extension of this approach for video retrieval. For this, we extract the key frames from video and use them to represent the research results of video as well as to do the relevance feedback.

Keywords: *Image and Video Indexing and Retrieval, Visualization, Relevance Feedback*

Tóm tắt:

Tìm kiếm video theo ngữ nghĩa vẫn còn là một vấn đề mở. Trong khi có rất nhiều nghiên cứu về phân tích nội dung của video, chỉ một số rất ít các nghiên cứu về biểu diễn kết quả tìm kiếm và tương tác với người sử dụng. Trong bài báo này, đầu tiên chúng tôi sẽ đề xuất giao diện đồ hoạ 2 chiều phù hợp cho hệ thống tìm kiếm ảnh, cho phép giao tiếp theo hai chiều : chiều từ hệ thống tới người sử dụng cho phép hệ thống biểu diễn kết quả tìm kiếm hiện thời đối với người sử dụng và chiều từ người sử dụng đến hệ thống cho phép người sử dụng cung cấp các thông tin phản hội lại hệ thống. Trong giao diện này, ảnh truy vấn sẽ nằm ở giữa trục toạ độ, các ảnh kết quả sẽ nằm trên hệ trục toạ độ 2 chiều với các biên độ là các độ đo tương tự theo các đặc trưng của ảnh kết quả và ảnh truy vấn. Một giải thuật cho phép hệ thống học các thông tin phản hồi từ người sử dụng cũng được đề xuất. Giải thuật này được cài đặt và thử nghiệm trên nhiều cơ sở dữ liệu khác nhau. Sau đó, một khả năng mở rộng cách tiếp cận này cho tìm kiếm video cũng được trình bày. Các khung hình chính được trích chọn từ video và được sử dụng để tìm kiếm cũng như tương tác với người sử dụng.

Từ khoá: *Chỉ số và tìm kiếm ảnh và video, biểu diễn kết quả, phản hồi người sử dụng*

1. INTRODUCTION

Image indexing and retrieval appeared from 1992 [Kato92]. Up to now, some results has been obtained but it's still an open problem in comparing with obtained results in text retrieval. It's still true that concerns the semantic image indexing and retrieval. Indeed, the images in image databases were indexed by their low features but not by their semantic contents. Therefore, it has always a semantic gap [Smeulders00].

In order to fill up this semantic gap, two main approaches are possible: one is to index the images by their semantic contents in the form of ontology [Maillot05], another that we will present in this paper is to interact with the users in order to understand their needs [Rui98].

Video indexing and retrieval was born later than image indexing and retrieval. With the video, we have other information that is the motion. In some cases, this information will help to fill up the semantic gap. The problem of semantic video indexing and retrieval is considered as video event recognition. Therefore, the semantic video indexing and retrieval has the results for some specific domains. However, we believe that with general video databases, the participation of the users in the form of relevance feedback is obligatory. Unfortunately, nowadays, we do not have yet any work dedicated for it because of its difficulty.

Therefore, in this article, firstly, we propose a 2D graphic interface adapted to the problem of image retrieval that enable a bidirectional communication: from the system towards the user to visualize the current research results and from the user towards the system so that the user can provide some relevance feedback information to refine his/her query. In this interface, the visualization shows the image query in the middle of the screen and the result images in a 2D plan with distances showing the similarity measures between images and the query. We propose also a method of relevance feedback in form of validation, in this interface, for image retrieval. This approach has been implemented and tested with different image databases. Secondly, we analyze the extension of this approach for video retrieval. For this, we extract the key frames from video and use them to represent the research results of video as well as to do the relevance feedback

The paper is organized as follows. In the next section, we review some related works in image and video indexing and retrieval. A proposed interface and a method of relevance feedback are presented in section 3. Some experiments and results are introduced in section 4. Finally, for concluding this paper, we give some conclusions and perspectives.

2. RELATED WORKS

In this paper, we are interested in the second approach in two above presented approaches. This approach has two challenges:

- The improvement the kernel of system (methods of relevance feedback) for corresponding better the user's need
- The visualization that allows the system to present the results to the user and to ask him/her the feedback information

Concerning to the first challenge, a lot of works have been carried out. Those works are either to modify the query and the used similarity measure [Rui98] or to try to approximate the decision surface that separates the relevant images and the non relevant images in the feature's space [Zhang05]. Concerning to the second approach, the traditional interface that shows the image query and a fixed number of image results in the form of a list, becomes unsuitable for interactive image indexing and

retrieval system. In the last years, the visualization has drawn the attention of many researchers in the field of image indexing and retrieval domain and some methods of visualization were proposed. In [Rubner98], the EMD (Earth Mover's Distance) and MDS (Multidimensional Scaling) are used to present the image results in a plan 2D according to only one similarity measure. This work lacks of the relevance feedback. In 2004, Deng [Deng04] has been proposed a method of visualization and comparison the images based on the SOM. At the beginning, this method does not support the relevance feedback. At present, some current works [Laaksonen05] try to add the relevance feedback in it.

In the video indexing and retrieval, many works dedicated for video indexing and retrieval from the approaches based on the motion and trajectory [Dagtas00] to the approaches, that are more semantic, based on event recognition [Calic05]. Because, with the video beside the visual features, we have also audio features and text features. Therefore, some authors consider also the multi modality mode [Babaguchi02] [Huang99]. Some researchers are interested in adding the relevance feedback in video indexing and retrieval. At present, some works [Yan03] [Aksoy05] are presented for video indexing and retrieval with the relevance feedback, but indeed, they work with the still images.

Our work is aim to solve two challenges in image indexing and retrieval in the same time by a proposed interface and a corresponding method of relevance feedback. We extend our interface to video indexing and retrieval based on their key frames.

3. PROPOSED INTERFACE AND RELEVANCE FEEDBACK

Figure 1 shows the architecture of our approach for the image and videos indexing and retrieval. The *distance computing, results displaying* and *feedback* processes will be described in section 3.1 and section 3.2. In this architecture, for working with video, at first, we select some key frames from the video and then use the same process of image indexing and retrieval. In the section 3.3, we will present some works for key frame selection.

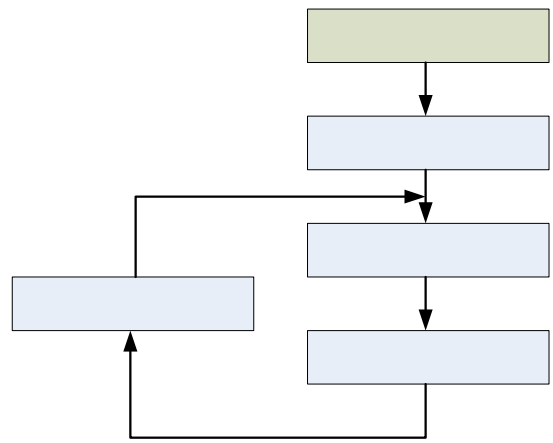

Figure 1. Architecture of proposed approach

3.1. Proposed interface

Our idea is to conceive an interface that allows to gather around the image query the images that are similar to. The user, then, select a region on the screen to refine his/her query. The figure 3 shows our proposed interface. This interface consists 5 zones: one of function, one of initial examples, one of positive examples, one of negative examples and another of results. The zone of result is the most important zone. In this zone, the image query is middle, the image results are shown according to 2 axes. We can use some different extracted features and their similarity measures to compute the position of result images in this plan. Without lost of generality, in this article, we present our

approach by using the histogram intersection of RGB color space for one axe and that of HSV color space for the other. The position (x, y) of image I in the image databases is computed as:

$$x(I) = d_{\text{Intersection}}(H_{RGB}(I), H_{RGB}(R))$$
$$y(I) = d_{\text{Intersection}}(H_{HSV}(I), H_{HSV}(R))$$
$$sign(x) = 1 \text{ if } E(H_{RGB}(I)) > E(H_{RGB}(R))$$
$$\text{and } sign(x) = -1 \text{ otherwise}$$
$$sign(y) = 1 \text{ if } E(H_{HSV}(I)) > E(H_{HSV}(R))$$
$$\text{and } sign(y) = -1 \text{ otherwise}$$

Where E is the entropy, H_{RGB}, H_{HSV} are the histogram in the RGB and HSV color spaces and $d_{Intersection}$ is the distance between two histograms [Swain91]. With this interface, we use only the global features. The local features might be added to improve the results.

3.2 Relevance feedback in the proposed interface

The presented method of relevance feedback in this section consist to do the relevance feedback by the modifying the query. This method works only with the positive examples in the form of validation. As soon as the current result are shown on the screen, the users can click on the images to select as positive examples, this image will be added in the zone of positive examples that is initialized by the image query.

The new query R' is set of n positive examples P_j. The new position (x',y') of image I is computed as:

$$x'(I) = \min(d_{\text{Intersection}}(H_{RGB}(I), H_{RGB}(P_j)))$$
$$\text{with } j \in (1,n)$$
$$y'(I) = \min(d_{\text{Intersection}}(H_{HSV}(I), H_{HSV}(P_j)))$$
$$\text{with } j \in (1,n)$$

We keep the value of sign(x) and sign(y) for sign(x') and sign(y'). The next results are shown according to their new positions. The users can do this process many times until he/she is satisfied.

3.3 Its extension to video retrieval

Temporal video segmentation is the first step towards video retrieval. Its goal is to divide the video stream into a set of meaningful and manageable segments (shots) that are used as basic elements for indexing. A shot is defined as an unbroken sequence of frames taken from one camera. There are two basic types of shot transitions: abrupt and gradual. Abrupt transitions (cuts) are simpler, they occur in a single frame when stopping and restarting the camera. Although many kinds of cinematic effects could be applied to artificially combine two shots, and thus to create gradual transitions, most often fades and dissolves are used. More than eight years of temporal video segmentation research have resulted in a great variety of algorithms. Early work focuses on cut detection, while more recent techniques deal with the harder problem - gradual transitions detection. A full overview of temporal video segmentation will be found in [Koprinska01]. After shots are segmented, the key frames that represent the salient content of the shot will be extracted. Depending on the content's complexity of the shot, one or more key frames can be extracted from a single shot. Because of its importance, much research effort has been given in key frame extraction. Progress has been made in this area, however, the existing approaches either are computationally expensive or can not effectively capture the major visual content. In [Zhuang98], the authors proposed a clustering based approach which is both efficient and effective. In [Wolf96], the authors proposed a new algorithm for identifying the key frames by using the motion information. Since effective temporal segmentation techniques and key frames

extraction techniques exist in the literature, we will use these techniques in our system.

4. EXPERIMENTAL RESULTS

4.1 Image/video databases

In order to evaluate our system, we used a subset of Corel's image databases (http://wang.ist.psu.edu/docs/related/). This image databases consist of 1000 images divided into 10 classes. The image size is either 384 × 256 or 256× 384. Figure 2 gives some images in this image database.

Figure 2. Some images in subset of COREL's image database

We have also tested our system with a video database. We have chosen CAVIAR video database [Fisher04]. This video database consists of two sets. The first set contained 28 videos were filmed for the CAVIAR project with a wide angle camera lens in the entrance lobby of the INRIA Labs at Grenoble, France. The resolution is half-resolution PAL standard (384×288 pixels, 25 frames per second) and compressed using MPEG2. The second set of data contained 52 videos also used a wide angle lens along and across the hallway in a shopping center in Lisbon. For each sequence, there are two time synchronized videos, one with the view across and the other along the hallway. The resolution is half-resolution PAL standard (384×288 pixels, 25 frames per second) and compressed using MPEG2. This video database is used in video surveillance.

4.2 Some results with proposed interface

4.2.1 Some results with image database

Figure 3 and figure 4 give a result of our proposed system with the corresponding method of relevance feedback. In this case, the users search the images containing one (some) horse(s). The figure 3 shows the results without the relevance feedback, while figure 4 shows the new results in this case, the users choose 4 positive examples. The new results show an improvement after one time of relevance feedback. Most images around the image query contain one or some horses. Figure 5 and figure 6 give the room of centered region in two cases.

Figure 3. Results without relevance feedback, image query is center of the screen, image results are shown along to two axes: the vertical axe using the intersection histogram in RGB color space, the horizontal axe using the intersection histogram in HSV color space.

Figure 4. Results after one time of relevance feedback, four positive examples have been chosen

Figure 5. A zoom of the centered region with the obtained results in the case without relevance feedback

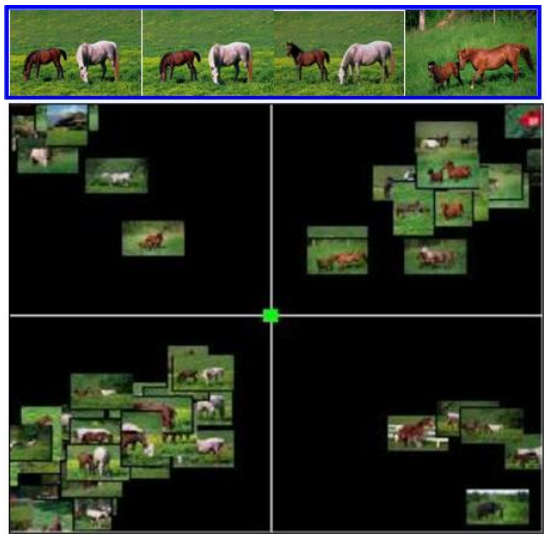

Figure 6. A zoom of the centered region with the obtained results in the case using one time of relevance feedback

4.2.2 Some results with video databases

For using proposed interface to present and retrieve the videos, some key frames must be extracted from these videos. For simplifying this task, with each video, we extract only one key frame. We propose here a method of extracting the key frames that is suited for this video database, based on the histogram intersection [Swain91]. We compare all of frame in the video with the first frame and choose the frame that has the largest distance with the first frame, as the key frame. In this paper, we compute histogram 24 bin of RGB color space (8 bins for each component). The figure 7 shows the key frame of scenario *Meet Crowd*. The figure 8 gives one result of the extension of our proposed interface for video indexing and retrieval based on the key frames.

Figure 7. An extracted key frame for video of scenario *Meet Crowd*

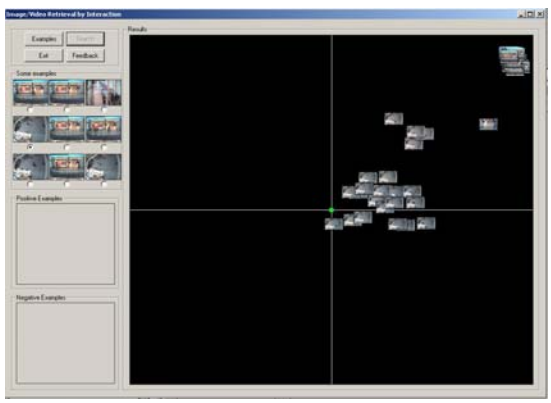

Figure 8. An extension of our proposed interface for video indexing and retrieval

4.3 Some comparisons

In order to compare our approach (proposed interface + relevance feedback) with others approaches (traditional interface + relevance feedback), we have also developed the proposed algorithm [Rui98] in the traditional interface. The figure 9 shows the first sixteen images results without relevance feedback. In this figure, the images in rectangle red are irrelevant, the others are relevant. The figure 10 shows the first sixteen images results after one time of relevance feedback.

Figure 9. The first sixteen images results without relevance feedback, the images in rectangle red are irrelevant, the others are relevant

Figure 10. The first sixteen images results after one time of relevance feedback

From the observations of obtained results, we give here some comparisons:

- The traditional interface presents the results in the mono dimensional way while our proposed interface presents these images in the bi-dimensional way. Therefore, it allows to show more information than the traditional interface. Our proposed interface gives implicitly the similarity between some groups of images (the traditional interface gives only the similarity between image query and the results images);

- Our interface allows to group the images that are similar (ex: in the figure 4 the group of flowers is on the right and below);

- The traditional interface limits the number of show image results, it is difficult for the users to give enough negative and positive examples. Our interface can solve this problem.

5. CONCLUSIONS

In this paper, a graphic 2D interface and a corresponding method of relevance feedback were presented. The presented experimental results of this interface and method of relevance feedback have been proved the efficiency. An extension to video indexing and retrieval is also introduced. However, there are some problems to solve for this interface such as: display the images in this interface when the image databases are large, relevance feedback with negative examples. In this paper, we have implemented an 2D interface and have tested only with histogram in RGB and HSV color spaces, this doesn't mean that we use only two features. With the images, we have more than 2 features, therefore, we are looking to two directions: one is to manipulate with an space more than 2 dimensions, another is to use the principal component analysis to reduce the number of dimensions. With the video indexing and retrieval, we consider also an interface that enables to present not only the key frames of the videos but also other information such as the motion and trajectory.

Reference

[Aksoy05] S. Aksoy, O. Cavus: "A Relevance Feedback Technique for Multimodal Retrieval of News Videos", in *Proceedings of EUROCON*, 2005.

[Babaguchi02] N. Babaguchi, Y. Kawai, T. Kitahashi: "Event-based indexing of broadcasted sports video by intermodal collaboration", *IEEE Trans. Multimedia*, vol. 4, 2002, p. 68-75.

[Calic05] J. Calic, N. Campbel, A. Calway, M. Mirhehdi, B. T. Thomas, T. Burghardt, S. Hannuna, C. Kong, S. Porter, N. Canagarajah, D. Bull: "Towards Intelligent Content Based Retrieval of Wildlife Videos", *Proc. to the 6th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS*, 2005.

[Dagtas00] S. Dagtas, W. Al-khatib, A. Ghafoor, R.L. Kashyap: "Models for motion based video indexing and retrieval", *IEEE Transactions on Image Processing*, vol. 9, num. 1, 2000.

[Deng04] D. Deng, J. P. M. Zhang: "Visualization and Comparison of Image Collections based on Self-organized Maps", *Proc. Information security, Data Mining and Web Intelligence, and Software*, 2004.

[Fisher04] R. B. Fisher: "The PETS04 Surveillance Ground-Truth Data Sets", *Proc. Sixth IEEE Int. Work. on Performance Evaluation of Tracking and Surveillance (PETS04)*, 2004, p.1-5.

[Huang99] J. Huang, Z. Liu, Y. Wang, Y. Chen, E. K. Wong: "Integration of multimodal features for video scene classification based on HMM", *in Proc. IEEE Workshop Multimedia Signal Processing*, Copenhagen, Denmark, 1999.

[Kato92] T. Kato, K. Hirata: "Query by visual example in content-based image retrieval", *Proc. EDB192. Lecture Notes in computer Science*, 1992, p. 56-71.

[Koprinska01] I. Koprinska, S. Carrato: "Temporal Video Segmentation: A survey", *Signal Processing: Image Communication*, 2001, p. 451-460.

[Laaksonen05] J. Laaksonen, V. Viitaniemi, M. Koskela: "Emergence of semantic concepts in visual databases", *In Proceedings of International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR 05)*, 2005, p. 127-134.

[Maillot05] N. Maillot: "Ontology Based Object Learning and Recognition", *PhD thesis, Université de Nice Sophia Antipolis*, 2005.

[Rubner98] Y. Rubner, C. Tomasi, L. J. Guibas: "A Metric for Distributions with Applications to Image Databases", *Proceedings of the Sixth International Conference on Computer Vision*, 1998.

[Rui98] Y. Rui, T. S. Huang, M. Ortega, S. Mehrotra: "Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval", *IEEE Transaction On Circuits and Video Technology*, 1998.

[Smeulders00] A. W. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain: "Content-Based Image Retrieval at the End of the Early Years", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, 2000, p. 1349-1380.

[Swain91] M. Swain, D. Ballard: "Color indexing", *International Journal of Computer Vision*, vol. 7, num. 1, 1991, p. 11-22.

[Wolf96] W. Wolf: "Key Frame Selection by Motion Analysis", *in Proceedings ICASSP'96*, 1996.

[Yan03] R. Yan, A. G. Hauptmann, R. Jin: "Negative Pseudo Relevance Feedback in Content based Video Retrieval", *ACM Multimedia 2003*, 2003.

[Zhang05] C. Zhang, X. Chen: "Region-Based Image Clustering and Retrieval Using Multiple Instance Learning", *CIVR*, 2005, p. 194-204.

[Zhuang98] Y. Zhuang, Y. Rui, T. S. Huang, S. Mehrotra: "Adaptive key frame extraction using unsupervised clustering", *in Proceedings of IEEE Int'l Conference on Image Processing*, 1998.

Authors:

Le Thi Lan graduated in Information Technology from Hanoi University of Technology. She has MS degree in Signal Processing and Communication from Hanoi University of Technology. She is currently PhD student of Project ORION, INRIA, France and Center MICA, Hanoi University of Technology, Vietnam. Her main research interests are in content-based indexing and retrieval, video understanding.

E-mail:Thi-Lan.Le@mica.edu.vn

Alain Boucher graduated in computer engineering from the Ecole Polytechnique of Montreal (Canada) in 1994. He received his Ph.D. degree from the Joseph Fourier University (Grenoble, France) in 1999. He worked on a European research project (ASTHMA) for 3 years at INRIA Sophia-Antipolis (France) until 2002. He is currently professor at the Francophone Institute for Computer Science (IFI-AUF) in Hanoi (Vietnam). His research interests include computer vision, content-based indexing and retrieval, pattern recognition and artificial intelligence.

E-mail: Alain.Boucher@auf.org

Monique Thonnat is a senior research scientist at INRIA (Director of Research 1rst class). She is the head of Orion a research group on cognitive vision at INRIA in Sophia Antipolis, France. She received in 1982 a PhD degree in Optics and Signal Processing from University of Marseille. In 1983 she joined INRIA in Sophia Antipolis. In 1991 she became Director of Research and in 1995 she created Orion, a multi-disciplinary research team at the frontier of computer vision, knowledge-based systems, and software engineering. Monique Thonnat is author or co-author of more than 100 scientific papers published in international journals or conferences; she has supervised 17 PhD theses. Her more recent research activities involve the conception of new techniques for the reuse of programs and on image understanding techniques for the interpretation of video sequences.

E-mail:
Monique.Thonnat@sophia.inria.fr