Special issue: The Future of Multimedia Analysis and Mining

Research Paper

Scalable approaches for content based video retrieval

Thanh DUC NGO[1], Duy DINH LE[2] and Shin'ichi SATOH[3]
[1]*The Graduate University for Advanced Studies (SOKENDAI),*
[2,3]*National Institute of Informatics*

ABSTRACT

This paper addresses content based video retrieval. First, we present an overview of a video retrieval framework and related approaches. Second, we consider two important applications of video retrieval nowadays which are video retrieval based on human face and video retrieval based on generic object categories. The goal is to develop approaches which require lowest annotation cost or computational cost while achieving competitive accuracy so that they can facilitate building scalable and comprehensive video retrieval systems.

KEYWORDS

Content-based video retrieval, scalable approaches, video mining

1 Introduction

Video retrieval refers to the task of retrieving the most relevant videos in a video collection, given a user query. A robust video retrieval system can bring benefits to a wide range of multimedia applications such as news video analysis, video-on-demand broadcasting, commercial video analysis, digital museums or video surveillance. In the past, when video collections are relatively small, video retrieval can be done using keywords manually annotated by specialist. However, due to recent exponential growth of video data supported by advances in multimedia technology, manual annotation has been no longer tractable. Consequently, it creates a great demand on automatic video retrieval systems.

In general, a video itself contains multiple types of information including embedded video metadata (e.g. title, description, creation date, author, copyright, duration, video format), audio content, and visual content. In the context of this paper, we address video retrieval systems based on information derived from visual content only.

2 Content based video retrieval

Building a video retrieval system requires solutions to several problems. We briefly introduce parts of a typical content based video retrieval framework (summarized in Fig. 1) in the following.

2.1 Video parsing

Videos are usually organized as a hierarchical structure of scenes, shots, and frames (illustrated in Fig. 2). The goal of video parsing is to divide a video into a set of such structural elements. Depending on application, video elements of a corresponding type will be used as the fundamental processing units. For instance, object based video retrieval may need to analyze videos at frame level. Meanwhile, event based video retrieval mainly targets to shots.

Video parsing is a prerequisite step towards video content analysis and indexing. Approaches for video parsing includes scene segmentation, shot boundary detection, and keyframe selection.

Shot boundary detection. A shot is defined as a sequence of frames captured by a single camera operation. The interruptions between camera operations indicate the shot boundaries, thus make frames in a shot strongly correlated each other. There are two basic categories of shot boundaries, depending on the transitions between shots. A shot boundary is categorized as CUT if the transition between shots is abrupt. An abrupt transition occurs in a single frame only. Otherwise, if a transition spreads over a number of frames, the

Received November 7, 2013; Accepted November 29, 2013.
[1)]*ndthanh@nii.ac.jp*, [2)]*ledduy@nii.ac.jp*, [3)]*satoh@nii.ac.jp*
DOI: 10.2201/NiiPi.2014.11.5

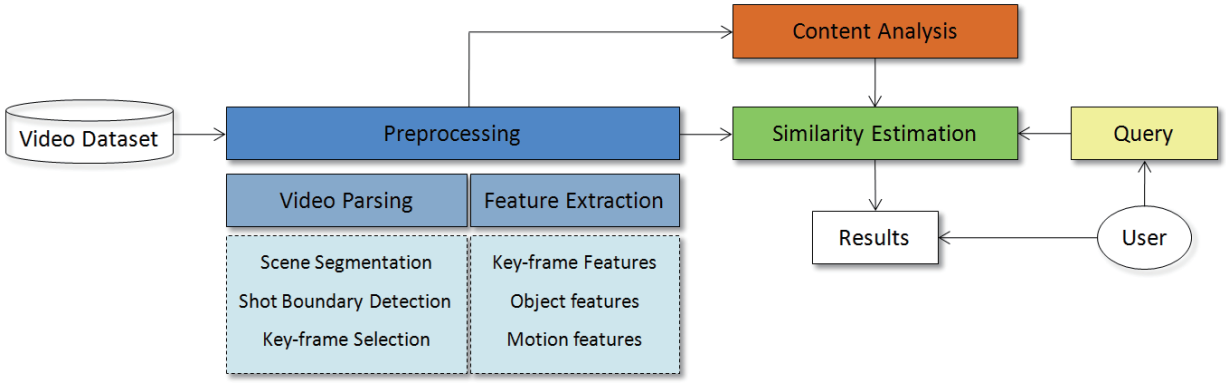


Fig. 1 An overview of a video retrieval system.

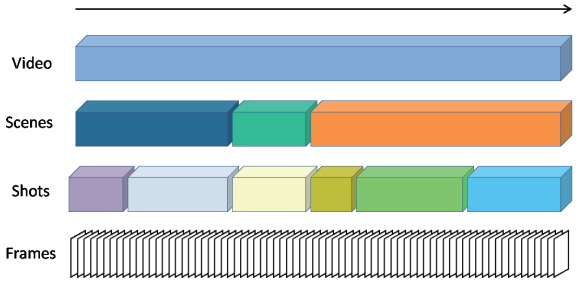


Fig. 2 An illustration of video structure.

shot boundary is called gradual transition (GT). Gradual transitions are mainly created by editing effects. Shot boundary detection (SBD) aims at detecting such transitions between consecutive shots.

Shot boundary detection approaches are usually based on measuring the dissimilarities between frames of which visual features are extracted. The dissimilarities between pairs of consecutive frames or between frames within a window [2] can be measured using several types of distance such as 1-norm cosine distance, Euclidean distance, histogram intersection distance, the chi squared distance [2], [3], or the Earth Mover's distance [1]. Using the measured dissimilarities, shot boundaries are detected by either thresholding [4], [5], graph partitioning [19], or applying learned classifiers.

Keyframe selection. Consecutive frames within a sequence (i.e. a video or a shot) are highly redundant. Thus, a set of certain frames which best reflect content of the sequence should be selected to represent the sequence. Such frames are called keyframes or representative frames. The ultimate goal of keyframe selection is to eliminate redundant frames while reserving salient frames as much as possible.

Recent approaches to keyframe selection mainly target to minimize the dissimilarities between each selected keyframe with its neighboring frames, or maximizing dissimilarities between the selected frames. Approaches of the first strategy include clustering based and curve simplification based approaches. On the other hand, approaches following the second strategy consist of those based on sequentially selecting a keyframe which is significant different to the previous selected keyframe [6], or minimizing the correlations between keyframes within the selected set.

Scene segmentation. A video usually consist of scenes, where each scene may contain one or more shots. Shots of a scene are about the same subject or theme. Thus, scenes are also known as story units and they are at higher semantic level than shots. Scene segmentation is to decompose a video into scenes. Regarding to [18], scene segmentation approaches can be divided into four categories mostly depending on their strategy such as merging, splitting, statistical modeling, and boundary classification.

Merging based approaches gradually merge similar shots to form a scene following a bottom up style [8]. In contrast, splitting based approaches split the whole video into separate coherent scenes using a top down style [9]. With approaches based on statistical model, they aim at constructing statistical models of shots such as stochastic Monte Carlo [7], GMM [10], or a unified energy minimization framework [12] for scene segmentation. With boundary classification based approach, they extract features of shot boundaries and then use them to classify the shot boundaries as scene or non scene boundaries [11].

2.2 Feature extraction

Given video elements parsed by video parsing approaches, the next crucial step to construct a video retrieval system is to extract features from the video ele-

ments so that they can be used for video content analysis. Common features include features extracted from static keyframes and motion features extracted from sequences of frames.

Static keyframes are the basic video elements reflecting video content. Feature of static keyframes are basically derived from colors, textures, and shapes in keyframes or their regions. Recent works usually employ Bag-of-Visual-Word (BoVW) model borrowed from text retrieval for feature presentation. *Visual words* (i.e. salient regions) are first extracted from keyframes. They are then used to compose histograms of *visual words* representing individual keyframes or shots. Because such features are extracted from static keyframes, they can not capture motions which are mainly caused by camera movements and foreground object movements in videos. Motion features play a significant role in video indexing and retrieval by events or human actions.

2.3 Content analysis

The aim of content analysis is to analyze videos so that they can be indexed and retrieved by their content. The indexed terms are of several types including common patterns, video genres [13], [14], events [15], human actions, object categories or concepts appearing in the videos [16], [27]. Video content analysis requires techniques for mining patterns of interest in videos, video classification and annotation.

Recent approaches for video classification and annotation follow a typical strategy. First, low level features are extracted. Then, related category or concept classifiers are trained. Finally, the classifiers are used to map the features of video elements (e.g., videos, shots, frames) to the corresponding labels of the concepts or categories. Basically, the main challenge is to handle the *semantic gap* between low level feature extracted from videos and semantic concepts perceived by human being (e.g., video genres, events, object categories). However, it is well known that bridging the semantic gap is challenging due the variations in visual appearance of semantic concepts. Furthermore, human participation, in the form of manual annotation, is always required in order to train classifiers. This makes video annotation and classification approaches inflexible as they are applied to different domains.

2.4 Query formation

Query formation is at the online stage of a video retrieval system. As a query is given, the retrieval system perform retrieval by applying similarity estimation approaches or simply scanning over an index table to return most relevance video elements in accordance with the query.

To formulate a query, users usually submit an example which visually represents what they want to search. This type of query is called query-by-example. Depending on the application as well as users' interest, the example can be a whole image, a bounding region of object of interest in an image, or a sketch [17]. Query-by-example is very useful when users want to search for the same object or scene under slightly varying circumstances and when the example images are available indeed. If proper example images are not available, it is impossible to perform searching. Query-by-example is considered as non semantic based video query type since it can not capture the semantic search intention from the example. Once again, this is due to the *semantic gap*. To narrow such a semantic gap, one way is to use textual keywords in queries which describe what the user wants to search. However, this would require a textual annotation of the retrieved video dataset. Manually annotating the dataset is extremely expensive.

Query-by-concept paradigm is yet another way to bridge the semantic gap between users' search intention and visual video content. With this paradigm, users can select a predefined concept, after which the retrieval system return relevant video elements based on presence of the concept detected by concept detectors or video annotation and classification approaches. By doing this, the semantic consistence between users' search intention and visual content of videos is kept. Furthermore, it bypasses the limitation of query-by-example paradigm such as needed existence of example image. However, because the concepts are predefined, the retrieval system cannot support searching concepts out of the cope. And, human participation is required in order to train the detectors.

2.5 Similarity estimation

Video similarity estimation play an important role in a content based video retrieval system. The choice of approaches depends on the query type.

Feature-based similarity estimation. This is mostly for query-by-example paradigm. The most direct measure of similarity between video elements and a query is the distance between their extracted features. According to different user' demands, features of static keyframes, object features can be used to measure their similarity. However, selecting appropriate types of feature is one of the most critical problems. Furthermore, the estimation process is costly and time consuming if the dataset is huge.

Concept-based similarity estimation. Matching the name of each concept with query terms is the simplest way of finding the videos that satisfy the query. Basically, if the concept detection is done for all videos of the retrieved dataset in the offline stage,

the retrieval system can respond to users' search request in constant time by scanning an inverted index table. If users retrieve multiple concepts simultaneously, returned videos elements can be ranked by voting. The limitation of this approach is that it only supports searching certain concepts with corresponding trained classifiers in advance.

3 Motivations and addressed problems

Firstly, scalability is no longer a plus feature but a definite requirement of nowadays video retrieval systems due to the exponential growth of video data. Applications involving video retrieval can hardly be practical if the system is not scalable. Investigating scalable approaches is therefore importance to content-based video retrieval.

A system is defined to be scalable if it can be easily extended to handle a much larger amount of data and its overall consumption of resources increase gracefully with the size of the database. In other words, the key factor that affects scalability of a system is its cost-effectiveness. There are two main types of cost possibly consumed by a content-based video retrieval systems: computational cost and human annotation cost. Human annotation cost can be regarded as the amount of manual annotation needed in developing the system such as annotation for training classifiers in content analysis approaches. Minimizing these costs is essential to achieve scalable retrieval systems. In accordance with that, following issues come into being: 1) how to reduce the costs in most expensive processes e.g. similarity estimation and content analysis ? and 2) how to balance cost-effectiveness and accuracy of a content-based video retrieval system while cost-effectiveness is usually inversely proportional to accuracy ? These issues must be considered in developing scalable retrieval system.

Secondly, despite a great deal of progress has been made in some of the core aspects of video retrieval, there is still much more room for improvement especially when scalability is taken into account. In the following, we present our scalable approaches for video retrieval based on matching faces and object categorization.

4 Face retrieval in large-scale datasets

Building a robust video retrieval system based on face is not a trivial task because of the fact that the imaged appearance of a face changes dramatically under large variations in poses, facial expressions, and complex capturing conditions. Moreover, efficiency is also an issue in such a face retrieval system because the scales of available datasets are rapidly getting larger, for instance, exceeding thousands of hours of videos with millions of faces belonging to hundreds of characters.

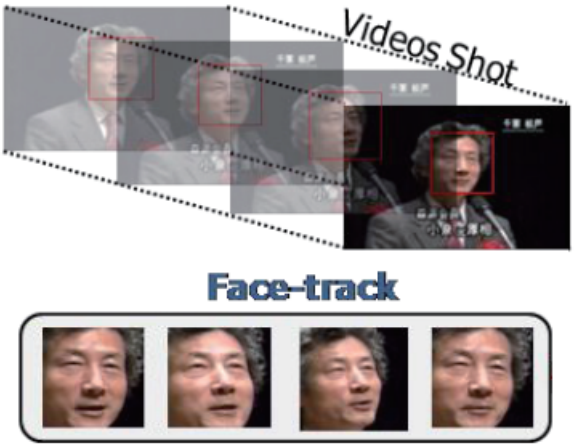


Fig. 3 Faces in a face-track with different facial expressions and poses.

A face retrieval system generally consists of two main steps. The first step is extracting the appearance of faces in videos. The second step is matching the extracted ones with a given query to return a ranked list. Whereas conventional approaches take into consideration single face images as the basic units in extracting and matching [20]–[22], recently proposed approaches have shifted toward the use of sets of face images called face-tracks. A face-track contains multiple face images belonging to the same individual character within a video shot. The face images in a face-track may present the corresponding character from different viewpoints and with different facial expressions (as shown in Fig. 3). By exploiting the plenteous information from the multiple exemplar faces in the face-tracks, face-track-based approaches are expected to achieve a more robust and stable performance. In this work, we addressed scalable approaches for both face-track extraction and face-track matching.

4.1 Face-track extraction

We propose a point tracker based face-track extraction approach, which is very efficient compared to approaches using an affine covariance region tracker or face clustering. The basic idea is that if two faces detected in different frames share a large amount of similar point tracks (i.e. trajectories of tracked points) passing through both of them, they are likely to be faces of the same character.

Assuming some points are generated and tracked through frames of a shot, we have the output of the tracking process as a set of tracking trajectories. One trajectory is for one generated point. We call such tra-

Table 1 Performance of the evaluated approaches.

Approaches	#total extracted FT
Everingham et al.	613/755 (81.19%)
Ours.	711/755 (94.17%)

jectories point tracks. Given two faces A and B in different frames and the set of point tracks, there are four types of point tracks regarding their intersection with the faces: (a) point tracks that pass through both A and B, (b) point tracks that pass through A but not B, (c) point tracks that pass through B but not A, and finally, (d) point tracks that do not pass through either A or B. A point track passes through a face if its point lies within the face bounding box in the corresponding frame.

A confidence grouping measure (CGM) that the two faces A and B belong to the same character can then be defined as:

$$CGM(A, B) = \frac{N_a}{N_b + N_c} \quad (1)$$

where N_a, N_b, and N_c are the number of tracks of types (a), (b), and (c). If $CGM(A, B)$ is larger or equal to a certain threshold, the two faces, A and B, are grouped into one face-track.

To make point tracks reliable and sufficient in number for grouping faces of multiple characters throughout a shot, we introduce techniques to handle problems due to flash lights, partial occlusions, and scattered appearances of characters.

When flash lights occur in a frame, they significantly change the intensity of the frame. Thus, the tracker cannot track points properly. To handle such problems, frames containing flash lights should be removed (called flash-frames). We measure the luminosity of the frames in the video shot. If the luminosity of a frame is significantly increased compared with its neighbors, the frame is declared to be a flash-frame and removed.

On the other hand, to handle partial occlusion and scattered appearances of characters, we detect and remove incorrectly tracked points whose tracks only pass through one of the two faces. Additional points are generated to replace those that have been removed.

We test our proposed approach for face-track extraction on 8 video sequences from different video broadcasting stations including NHK News 7, ABC News, and CNN News. We directly compare our approach with the state-of-the-art approach proposed by Everingham et al. [23] in our experiment.

Experimental results (shown in Table 1) indicate that our proposed techniques and solutions are robust and efficient enough for extracting face-tracks in real-world news videos by successfully extracting 94% of all face-tracks. Our approach outperforms the approach in [23]. In terms of speed, our approach is approximately 2 times slower than that of Everingham et al. However, our complexity is somehow linear to the total number of faces. Meanwhile, Everingham et al. compared all pairs of faces in the shot. Their complexity is polynomial to the total number of faces. If the number of faces increases, the gap will be narrowed rapidly.

4.2 Face-track matching

Several approaches for matching face-tracks have been proposed. Although these approaches have shown high accuracy in benchmark datasets, their high computational costs limit their practical applications in large-scale datasets. This motivates us to target a matching approach which is extremely efficient while achieving a competitive performance with state-of-the-art approaches.

To maintain competitive accuracy, we keep using multiple faces of a face-track. However, instead of using all the faces in a face-track, we propose to subsample faces of a face-track regarding their temporal order of appearance. The neighboring range is controlled by a variable, k. With a given value of k, our approach starts by temporally dividing the face-track into k equal parts. The middle face of each part is selected to represent all faces within the part. Given k selected faces from the face-track and their extracted facial features, the face-track becomes a set of k points distributed in a feature space. We compute the mean point to represent the set. The distance between two sets now relies on the distance between their mean points. In other words, if the mean point is called a mean face, the similarity between two face-tracks corresponds to the distance between their mean faces. We call our approach k-Faces.

We conduct experiments on two large-scale face-track datasets obtained from real-world news videos. One dataset contains 1,497 face-tracks of 41 characters extracted from 370 hours of TRECVID videos. The other dataset provides 5,567 face-tracks of 111 characters observed from a television news program (NHK News 7) over 11 years. All face-tracks are automatically extracted using our proposed face-track extraction approach. We make both datasets publically accessible [1]. A statistical overview of the datasets are given in Fig. 4. We compared k-Faces with several approaches, including those based on pair-wise distances (min-min), MSM [24], and CMSM [25].

The results (shown in Table 2) generally demonstrate that our proposed approach is extremely efficient while achieving performance comparable with that of state-

[1] http://satoh-lab.ex.nii.ac.jp/users/ndthanh/ftrack

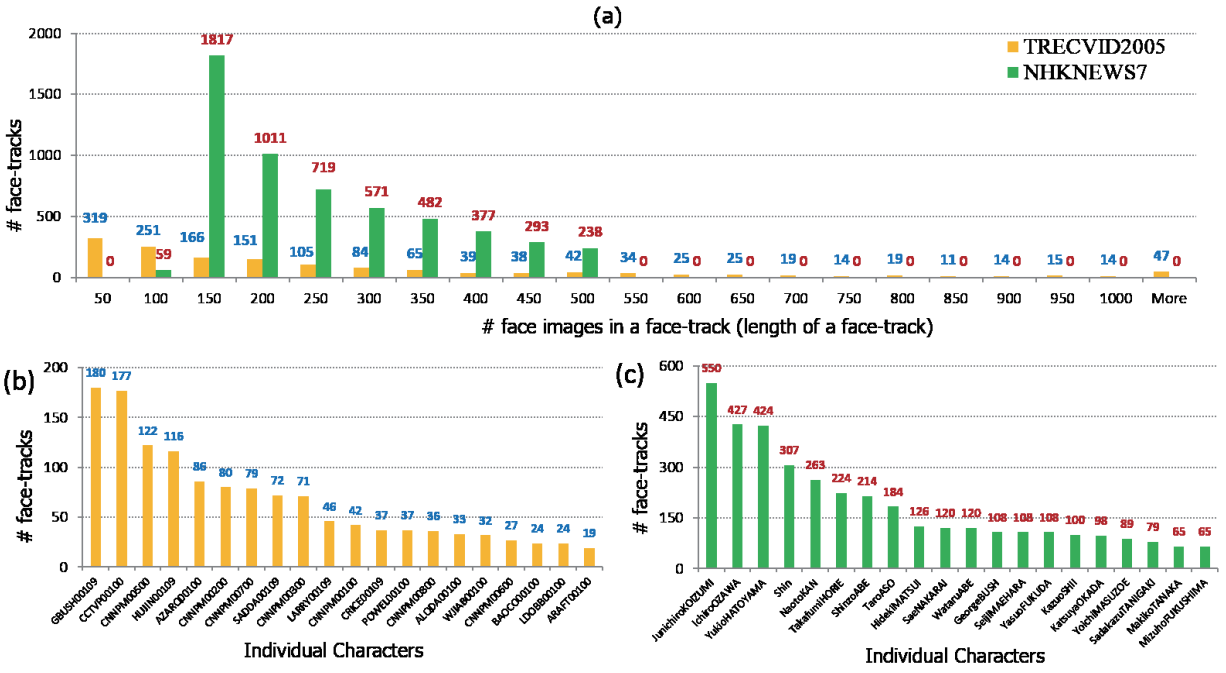


Fig. 4 Statistical overview of our face-track datasets. (a) shows the distribution of face-tracks over their lengths; (b) and (c) present the number of face-tracks for the top 20 individual characters.

Table 2 Mean Average Precision and processing times (in seconds) of the evaluated approaches.

Approaches	TRECVID		NHKNews7	
	MAP(%)	Matching Time	MAP(%)	Matching Time
pair-wise (min-min) + $L1$	76.54	2544.73	60.99	6678.00
k-Faces + $L1$ (k=20)	73.65	1.63	53.68	3.23
MSM	69.20	347.39	58.92	667.15
CMSM	64.62	95.36	53.08	155.40

of-the-art approaches. More details are presented in [28].

5 Object categorization for video retrieval

Video retrieval based-on concepts such as predefined object categories requires an object categorization approach which is to detect presences of the object categories in video at frame-level. Basically, such an approach require annotated data for learning object classifiers. To reduce annotation cost thus make the approach more scalable, annotations are given at image level instead of region level. Label of an image indicates whether the image contains an object but not its location. Low categorization accuracy may result since object region and background region within one image share the same label. To eliminate labeling ambiguity, we investigate Multiple Instance Learning (MIL) based object categorization approaches.

In MIL setting, groups and their samples are usually called bags and instances. A training group is labeled positive, if it has at least one positive instance. Otherwise, it is labeled as a negative bag. Given training labels are for groups, MIL approaches can learn to classify samples of the groups. If we consider a *bag* as *an image* (i.e., frame) and *instances in the bag* as *sub-windows* in the image, MIL approaches can be applied to detect object regions. However, one drawback of existing MIL approaches is that they disregard the correlations or spatial relations between sub-windows (i.e. instances) within an image (i.e. bag) in their iterative learning process. Such relations between sub-

windows can be clearly observed as: if a sub-windows is said containing an object, its highly overlapped sub-windows should contain the object also. We propose to improve the approaches by incorporating spatial information into learning. By doing so, the proposed approach improves categorization accuracy, compared to existing approaches. It therefore achieves a better balance between annotation cost and accuracy.

Formulated objective

Given a set of input instances $x_1, \ldots, x_n$ grouped into non-overlapping bags $B_1, \ldots, B_m$, with $B_I = \{x_i : i \in I\}$ and index sets $I \subseteq \{1, \ldots, n\}$. Each bag B_I is then given a label Y_I. Labels of bags are constrained to express the relation between bag and instances in the bag as follows: if $Y_I = 1$ then at least one instance $x_i \in B_I$ has label $y_i = 1$, otherwise, if $Y_I = -1$ then all instances $x_i \in B_I$ are negative: $y_i = -1$.

Let denote $x_{mm}(I)$ as the instance of bag B_I that has the highest positive score. And, $\mathcal{SR}(x_{mm}(I), T)$ is the set of $x_{mm}(I)$ and instances that surround $x_{mm}(I)$ with respect to the overlap parameter T. An instance belongs to $\mathcal{SR}(x_{mm}(I), T)$ if its overlap degree with $x_{mm}(I)$ is greater or equal to T, where $0 < T \leq 1$. The overlap degree between two instances (i.e. sub-windows) is the fraction of their overlap area over their union area. Then, our formulated objective can be presented as follows.

$$\begin{aligned} &\min_{\{y_i\}} \min_{\{w,b,\xi\}} \quad \frac{1}{2}\|w\|^2 + C\sum_I \xi_I \\ &\text{s.t.} \forall I: \; Y_I = -1 \wedge -\langle w, x_i\rangle - b \geq 1 - \xi_I \,, \; \forall i \in I \,, \\ &\quad \text{or} \quad Y_I = 1 \wedge \langle w, x^*\rangle + b \geq 1 - \xi_I \,, \\ &\quad \forall x^* \in \mathcal{SR}(x_{mm}(I), T) \,, 0 < T \leq 1, \text{and } \xi_I \geq 0 \end{aligned} \tag{2}$$

The constraints express that, if a sub-window in an image is classified as a positive instance, its neighboring sub-windows should be positive also. For example, if a sub-window tightly covers an object, its slightly surrounding sub-windows also contain that object.

Optimization by support vector machine

Our formulation can be cast as a mixed integer program in which the integer variables are the selectors of $x_{mm}(I)$ and the instances in $\mathcal{SR}(x_{mm}(I), T)$. This problem is hard to solve for the global optimum. However, we exploit the fact that if the integer variables are given, the problem reduces to a quadratic programming (QP) problem that can be completely solved. Based on that insight, we propose solution have an outer loop and an inner loop. The outer loop sets the values for the integer variables. Meanwhile, the inner loop trains a standard SVM. The outer loop stops if none of the integer variables changes in consecutive rounds. An illustration of a learning round is given in Fig. 5.

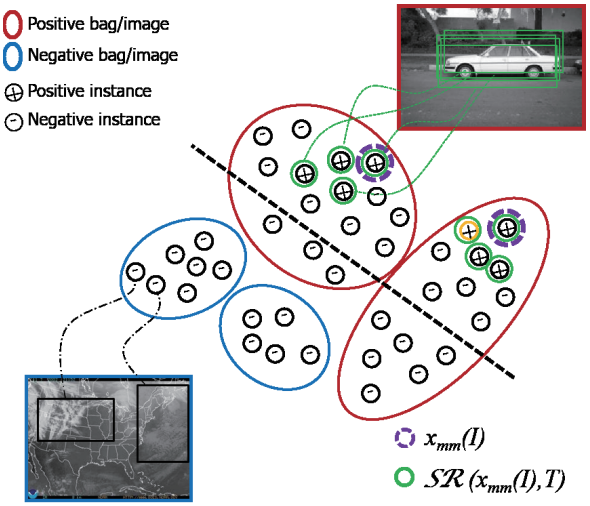


Fig. 5 An illustration of a learning round. The (black) dash line indicates the SVM hyper-plane at the current round. Relying on the hyper-plane, $x_{mm}(I)$ and its surrounding instances (on image space) in a bag are selected and grouped into $\mathcal{SR}(x_{mm}(I), T)$. All instances of $\mathcal{SR}(x_{mm}(I), T)$ are then treated as positive samples while other instances are treated as negative samples for training the next SVM classifier in the next round.

We then obtain the SVM classifier for sub-window (i.e. instances) classification. Given an unlabeled image (i.e. bag), the classifier can be used to classify the image by finding the sub-window that maximizes the score of the sub-window classifier. If this score is positive, the image is said to be positive, which means it contains the object of interest. In addition, the sub-window yielding the maximum score is the most representative region in the image for the presence of the object.

We perform experiments on benchmark Caltech-4 and Caltech-101 datasets. Images and sub-windows are presented by using standard Bag-of-Visual-Words model. We compare the proposed approach to the original SVM-based MIL approaches (mi-SVM and MI-SVM, introduced by Andrews et al. [26]) and two other standard approaches called GSC and MA. GSC and MA are SVM-based classifier trained at image level and sub-windows level respectively. Table 3 shows that our proposed approach outperforms other approaches by integrating spatial information into learning. In addition, we learned that accuracy can be improved up to 91.58% on Caltech-101 dataset if the proposed approach is combined with the GSC in a generalize stacking framework. The reason is because the combined classifiers complement each other. The GSC captures the global scene configuration in which an object ap-

Table 3 Average classification accuracy of the evaluated approaches.

Approaches	Accuracy (%)	
	Caltech-4	Caltech-101
MA	90.73	78.32
GSC	94.46	83.53
mi-SVM	72.54	60.49
MI-SVM	95.74	84.25
Ours	**96.28**	**87.26**

pear. Meanwhile, our approach is good at identifying local regions which best represent the object.

6 Conclusion

In this paper, we introduce approaches towards scalable content-based video retrieval systems. First, we address video retrieval based on human face. We presented robust and efficient approaches for face-track extraction and face-track matching. The matching approach achieved competitive accuracy compared to state-of-the-art approaches while it is hundreds to thousands times faster. Second, we target video retrieval based on object categories appearing in videos. The goal is to improve accuracy with minimum annotation data required. We introduced an approach based on Multiple Instance Learning. Spatial information is taken into account to achieve a significant accuracy improvement. The proposed approaches are expected to be helpful in building scalable and comprehensive retrieval systems.

References

[1] C. H. Hoi, L. S. Wong, and A. Lyu, "Chinese university of Hong Kong at TRECVID 2006: Shot boundary detection and video search," In *Proc. TREC Video Retrieval Eval.*, 2006.

[2] M. Cooper, T. Liu, and E. Rieffel, "Video segmentation via temporal pattern classification," *IEEE Trans. Multimedia*, vol.9, no.3, pp.610–618, 2007.

[3] G. Camara-Chavez, F. Precioso, M. Cord, S. Phillip-Foliguet, and A. de A. Araujo, "Shot boundary detection by a hierarchical supervised approach," In *Proc. Int. Conf. Syst. Sig. Image Process.*, 2007.

[4] Choi, K.-C. Ko, Y.- M. Cheon, G-Y. Kim, H-Il, S.-Y. Shin, and Y.-W. Rhee, "Video shot boundary detection algorithm," *Comput. Vis. Graph. Image Process.*, (*Lect. Notes Comput. Sci.*), vol.4338, pp.388–396, 2006.

[5] Z. Cernekova, I. Pitas, and C. Nikou, "Information theory-based shot cut/fade detection and video summarization," *IEEE Trans. Circuits Syst. Vid. Tech.*, vol.16, no.1, pp.82–90, 2006.

[6] X.-D. Zhang, T.-Y. Liu, K.-T. Lo, and J. Feng, "Dynamic selection and effective compression of key frames for video abstraction," *Pattern Recognit. Lett.*, vol.24, no.9/10, pp.1523–1532, 2003.

[7] Y. Zhai and M. Shah, "Video scene segmentation using Markov chain Monte Carlo," *IEEE Trans. Multimedia*, vol.8, no.4, pp.686–697, 2006.

[8] Z. Rasheed and M. Shah, "Scene detection in Hollywood movies and TV shows," In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2003.

[9] Z. Rasheed and M. Shah, "Detection and representation of scenes in videos," *IEEE Trans. Multimedia*, vol.7, no.6, pp.1097–1105, 2005.

[10] Y.-P. Tan and H. Lu, "Model-based clustering and analysis of video scenes," In *Proc. IEEE Int. Conf. Image Process.*, 2002.

[11] N. Goela, K. Wilson, F. Niu, A. Divakaran, and I. Otsuka, "An SVM framework for genre-independent scene change detection," In *Proc. IEEE Int. Conf. Multimedia Expo.*, 2007.

[12] Z. W. Gu, T. Mei, X. S. Hua, X. Q. Wu, and S. P. Li, "EMS: Energy minimization based video scene segmentation," In *Proc. IEEE Int. Conf. Multimedia Expo.*, 2007.

[13] M. J. Roach, J. S. D. Mason, and M. Pawlewski, "Motion-based classification of cartoons," In *Proc. Int. Symp. Intell. Multimedia*, 2001.

[14] Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *IEEE Trans. Circuits Syst. Video Tech.*, vol.15, no.1, pp.52–64, 2005.

[15] L. Xie, Q. Wu, X. M. Chu, J. Wang, and P. Cao, "Traffic jam detection based on corner feature of background scene in video-based ITS," In *Proc. IEEE Int. Conf. Netw. Sens. Control*, 2008.

[16] C. S. Xu, J. J. Wang, H. Q. Lu, and Y. F. Zhang, "A novel framework for semantic annotation and personalized retrieval of sports video," *IEEE Trans. Multimedia*, vol.10, no.3, pp.421–436, 2008.

[17] W. M. Hu, D. Xie, Z. Y. Fu, W. R. Zeng, and S. Maybank, "Semantic based surveillance video retrieval," *IEEE Trans. Image Process.*, vol.16, no.4, pp.1168–1181, 2007.

[18] W. Hu, N. Xie, X. Zeng, and S. Maybank, "A Survey on Visual Content-based Video Indexing and Retrieval," *Trans. Sys. Man Cyber.*, vol.1, no.6, pp.797–819, 2011.

[19] J. Yuan, B. Zhang, and F. Lin, "Graph partition model for robust temporal data segmentation," In *Proc. PAKDD*, 2005.

[20] M. Turk and A. Pentland, "Face recognition using eigenfaces," In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 1991.

[21] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 1994.

[22] P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *Trans. Pattern Anal. Machine Intell.*, vol.19, no.7, pp.711–720, 1997.

[23] M. Everingham, J. Sivic, and A. Zisserman, "Taking the bite out of automated naming of characters in TV video," *Image Vis. Comp.*, vol.27, no.5, pp.545–559, 2009.

[24] O. Yamaguchi, K. Fukui, and K. Maeda, "Face recognition using temporal image sequence," In *Proc. IEEE Conf. Auto. Face Gesture Reg.*, 1998.

[25] K. Fukui and O. Yamaguchi, "Face recognition using multi-viewpoint patterns for robot vision," *Intl. Symp. Robot. Research*, 2003. The Eleventh International Symposium of Robotics Research, ISRR, October 19–22, 2003, Siena, Italy. Springer 2005, pp.192–201. Springer Tracts in Advanced Robotics ISBN 987-3-540-23214-8.

[26] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple-instance learning," In *Proc. Neural. Info. Proces. Sys.*, 2003.

[27] C. G. M. Snoek, M. Worring, J. C. van Gemert, J. M. Geusebroek, and A. W. M. Smeulders, "The challenge problem for automated detection of 101 semantic concepts in multimedia," In *Proc. ACM Int. Conf. Multimedia*, 2006.

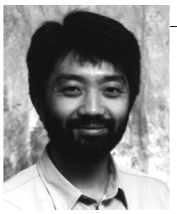
[28] T. D. Ngo, D.-D. Le, and S. Satoh, "Face retrieval in large-scale news video datasets," *IEICE Trans. Info. Sys.*, vol.E96-D, no.8, pp.1811–1825, 2013.

Thanh DUC NGO

Thanh DUC NGO obtained his BS degrees in Computer Science from the University of Science, Ho Chi Minh City, Vietnam in 2006. He pursuited his PhD degree at the Department of Informatics, The Graduate University for Advanced Studies (SOKENDAI), Japan. His research interests include pattern recognition, computer vision and multimedia analysis.

Duy DINH LE

Duy DINH LE received his BS and MS degrees in 1995 and 2001, respectively, from the University of Science, Ho Chi Minh City, Vietnam, and his PhD degree in 2006 from The Graduate University for Advanced Studies (SOKENDAI), Japan. He is currently an associate professor at the National Institute of Informatics (NII), Japan. His research interests include semantic concept detection, video analysis and indexing, pattern recognition, machine learning, and data mining.



Shin'ichi SATOH

Shin'ichi SATOH is a professor at the National Institute of Informatics (NII), Japan. He received his BE degree in 1987, ME and PhD degrees in 1989 and 1992 from the University of Tokyo. His research interests include video analysis and multimedia databases. He was a visiting scientist at the Robotics Institute, Carnegie Mellon University, from 1995 to 1997. He is a member of IPSJ, ITEJ, IEEE-CS, and ACM.