Video Retrieval by Managing Uncertainty in Concept Detection using Dempster–Shafer Theory

Kimiaki Shirahama
Graduate School of Economics, Kobe University
2-1, Rokkodai, Nada, Kobe 657-8501, Japan
shirahama@econ.kobe-u.ac.jp

Kenji Kumabuchi and Kuniaki Uehara
Graduate School of System Informatics, Kobe University
1-1, Rokkodai, Nada, Kobe 657-8501, Japan
kumabuchi@ai.cs.kobe-u.ac.jp, uehara@kobe-u.ac.jp

***Abstract*—This paper focuses on *concept-based* video retrieval which examines whether a shot is relevant or irrelevant to a query based on detection results of concepts, like *Person*, *Building* and *Car*. One key problem is *uncertainty* in concept detection. Even for state-of-the-art methods, it is difficult to accurately detect concepts present in a shot. Relying on such uncertain concept detection results degrades retrieval performance. To overcome this problem, *Dempster–Shafer Theory* (DST) is used to represent the probability that a concept is possibly present in a shot. By incorporating DST into maximum likelihood estimation, our method estimates the probabilistic distribution of concepts' presences which characterize shots relevant to the query. A preliminary experiment on TRECVID 2009 video data supports the effectiveness of our method.**

Keywords*-*Video retrieval; Concept Detection; Uncertainty; Dempster-Shafer Theory; Evidential EM Algorithm

I. INTRODUCTION

Video retrieval can be treated as a machine learning process, one that constructs a classifier for discriminating between shots that are relevant or irrelevant to a query. A large number of example shots are required to construct a classifier that can accurately retrieve relevant shots, irrespective of object appearance, environment and camera technique. However, it is impractical to prepare enough example shots to suit all possible queries. This insufficiency of example shots is a key factor in the challenging problem of the *semantic gap* between low-level features computed automatically and high-level semantics perceived by human.

To bridge the semantic gap, one promising approach is *concept-based* video retrieval which retrieves shots, where concepts (*e.g.*, *Person*, *Building* and *Car*) related to a query are detected. This approach utilizes *concept detectors* that detect the presence of a concept in a shot. These are constructed using a large number of training shots that are annotated to indicate the presence or absence of a concept. Hence, the concept can be detected robustly, irrespective of its size, position and direction on the screen. A large number of researchers reported that using such concept detection results as 'intermediate' features significantly improves retrieval performance [1], [2], [3].

Figure 1 outlines concept-based video retrieval. First of all, a shot is associated with *concept detection scores*, each of which represents the probability of a concept's presence (Figure 1 (d)). Given a query represented using text and example shots (*i.e.*, 'multimodal query' in Figure 1 (a)), concepts related to the query are selected (Figure 1 (b)). A classifier is then constructed to discriminate between relevant and irrelevant shots to the query, using detection scores of selected concepts (Figure 1 (c)).

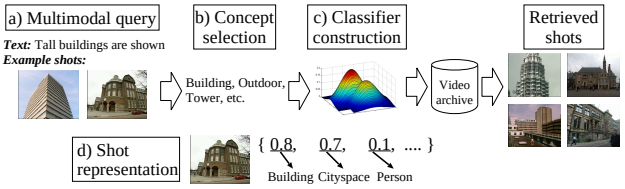

Figure 1. An overview of concept-based video retrieval.

We will now summarize the tasks necessary for concept-based video retrieval. The first is how to define a vocabulary of concepts. The most popular vocabulary is the Large-Scale Concept Ontology for Multimedia (LSCOM) [4]. LSCOM defines a standardized set of $1,000$ concepts in the broadcast news video domain. These are selected based on their 'utility' for classifying content in videos, their 'coverage' for responding to a variety of queries, their 'feasibility' for automatic detection, and the 'availability' (or 'observability') for large-sized training data.

The second task is how to select concepts related to a query. Several concept selection methods have been proposed so far. For example, concepts can be selected based on their lexical similarity to query terms and on detection scores in example shots [1], [7]. We have also developed a concept selection method using various concept relationships (*e.g.*, generalization/specialization, sibling, part-of *etc.*) defined in a knowledge base [8].

The last task that this paper addresses is how to construct a classifier that discriminates between relevant and irrelevant shots to a query. In this task, detection scores for multiple concepts are fused into a single *relevance score*, which represents the relevance of a shot to the query. However, even for most effective methods, it is difficult to accurately detect

 ISBN: 978-1-61208-195-3

any kind of concept. For example, TRECVID is an annual competition where concept detectors developed all over the world are benchmarked using large-scale video data [5]. At TRECVID 2010, the top-ranked methods achieved high performances for concepts such as *Mountain* and *Vehicle* (with average precisions greater than 0.2). On the other hand, the detection of concepts like *Bus* and *Sitting_down* was difficult (with average precisions less than 0.05). Thus, relying on such 'uncertain' concept detection results significantly degrades retrieval performance.

We introduce a method which constructs a classifier based on uncertain concept detection scores using *Dempster–Shafer Theory* (DST) [9]. DST is a generalization of Bayesian theory, where a probability is not assigned to a variable, but instead to a subset of variables. Such a probability is called a *belief mass*. We consider two variables P and A which represent the presence and absence of a concept in a shot, respectively. In addition, we consider $\{P, A\}$ which represents the uncertainty of whether the concept is present or not. Based on these variables, we define three belief masses $m(\{P\})$, $m(\{A\})$ and $m(\{P, A\})$. Here, $m(\{P\})$ and $m(\{A\})$ denote the probability that the concept is definitely present in a shot, and the probability that it is definitely absent, respectively, while $m(\{P, A\})$ denotes the probability that the concept is possibly present in the shot. By incorporating belief masses into maximum likelihood estimation, we can construct a classifier that can account for the uncertainty in concept detection.

II. Related Work

We will review existing methods for constructing classifiers based on concept detection scores. These classifiers can be roughly grouped into four categories: *linear combination*, *discriminative*, *similarity-based* and *probabilistic*. Linear combination classifiers compute the relevance score of a shot by weighting detection scores for multiple concepts. Popular weighting methods use the lexical similarity between query terms and a concept, their correlation (co-occurrence), and the detection scores of the concept in example shots [1], [7].

Discriminative classifiers consider a shot as a multi-dimensional vector, where each dimension represents the detection score of a concept. Based on this, a discriminative classifier, typically an SVM, is constructed using example shots [1], [3]. The relevance score of a shot is obtained as the classifier's output.

Similarity-based classifiers compute the relevance score of a shot as its similarity to example shots in terms of concept detection scores. Li *et al.* used the cosine similarity and a modified entropy as similarity measures [10].

Finally, probabilistic classifiers estimate a probabilistic distribution of concepts using concept detection scores in example shots, and use it to compute the relevance score of a shot. Rasiwasia *et al.* computed the relevance score as the similarity between the multinomial distribution of concepts estimated from example shots and the multinomial distribution estimated from the shot [11].

Our method constructs a classifier that is an extension of probabilistic classifiers. Specifically, ordinal probability (or Bayesian) theory cannot represent the uncertainty of a concept's presence in a shot. The only way to represent the uncertainty is to assign 0.5 to probabilities of the concept's presence and absence. Compared to this, DST can represent the uncertainty using $m(\{P, A\})$. Therefore, the representation of concept detection scores in our method is much more powerful than that of existing methods. To the best of our knowledge, such a representation has not been used in any previous methods.

III. Concept-based Video Retrieval based on Evidential EM Algorithm

In this section, we present a classifier construction method that accounts for the uncertainty in concept detection. First, we describe a method that computes the *plausibility* of a concept's presence (or absence) by combining belief masses for the concept. The plausibility represents the upper bound of probability that the concept is present (or absent) in a shot [9]. Thus, the plausibility of the concept's presence is useful for recovering false negative detections, while the plausibility of its absence is useful for alleviating false positive detections. We then present a probabilistic model based on plausibilities and *Evidential EM* ($\mathrm{E}^2\mathrm{M}$) algorithm, which estimates parameters of the model based on maximum likelihood estimation [9].

Plausibility computation based on DST: Let s_i^j be the detection score of the i-th shot ($1 \leq i \leq N$) for the j-th concept ($1 \leq j \leq M$). Based on s_i^j, we consider three belief masses $m_i^j(\{P\})$, $m_i^j(\{P, A\})$ and $m_i^j(\{A\})$, where the superscript j and the subscript i represent the j-th concept and the i-th shot, respectively. DST offers various combinations of belief masses based on set-theoretic operations. The following combination is used to compute the plausibilities pl_i^{j1} of the j-th concept's presence, and pl_i^{j0} of its absence:

$$\begin{aligned} pl_i^{j1} &= \sum_{B \cap \{P\} \neq \phi} m_i^j(B) = m_i^j(\{P\}) + m_i^j(\{P, A\}), \\ pl_i^{j0} &= \sum_{B \cap \{A\} \neq \phi} m_i^j(B) = m_i^j(\{A\}) + m_i^j(\{P, A\}) \end{aligned} \quad (1)$$

where the presence and absence of the j-th concept are represented by '1' and '0', respectively. B indicates any subset of variables overlapping $\{P\}$ or $\{A\}$. For pl_i^{j1}, $\{P\}$ and $\{P, A\}$ are referred by B as shown in the right-hand side. Thus, by defining pl_i^{j1} as the sum of $m_i^j(\{P\})$ and $m_i^j(\{P, A\})$, almost all shots where the j-th concept is present have relatively large pl_i^{j1}. The same is true of pl_i^{j0}.

To compute pl_i^{j1} and pl_i^{j0}, we extract three types of intervals using s_i^j. The first type characterizes $m_i^j(\{P\})$

where the number of shots annotated with the j-th concept's presence is much larger than the number of shots annotated with its absence. In the second type of interval characterizing $m_i^j(\{A\})$, the number of the latter type of shots is much larger than the number of the former type of shots. The last type of interval for $m_i^j(\{P,A\})$ does not have a distribution that is biased towards shots annotated with the j-th concept's presence or absence. However, directly estimating $m_i^j(\{P\})$, $m_i^j(\{A\})$ and $m_i^j(\{P,A\})$ is difficult, since we have no priori knowledge about their probabilistic distributions. Thus, following the spirit of DST in equation (1), we compute pl_i^{j0} and pl_i^{j1} based on the *lower and upper bounds* of s_i^j, which are defined as the minimum and maximum of the interval for $m_i^j(\{P,A\})$, respectively. Thereby, almost all shots where the j-th concept is presence have s_i^j larger than the lower bound, while almost all shots where it is absent have s_i^j smaller than the upper bound.

To implement the above idea, we construct a linear SVM using shots annotated with the j-th concept's presence and absence. In Figure 2, the former and latter cases are represented as $\times$s and $+$s, respectively, where each shot is represented using s_i^j (*i.e.*, a real number). As shown in Figure 2, we use the left and right support vectors as the lower and upper bounds. It can be considered that if s_i^j is larger than the lower bound, the probability of the j-th concept's presence in the i-th shot is at least greater than 0; that is, $pl_i^{j1} > 0$. It is also reasonable to assume that a larger pl_i^{j1} is computed for a larger s_i^j. Hence, pl_i^{j1} is computed using *Line 1* in Figure 2, where pl_i^{j1} is 0 at the lower bound and 1 at $s_i^j = 1$. Similarly, *Line 0* is used to compute pl_i^{j0}, where pl_i^{j0} is 0 at the upper bound and $1/\rho$ at $s_i^j = 0$.

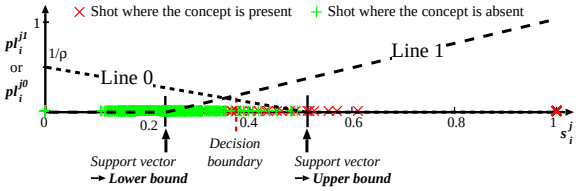

Figure 2. Illustration of the plausibility computation using support vectors.

The following two points are important for the computation of pl_i^{j1} and pl_i^{j0}. The first is that, in order for the interval between the lower and upper bounds to include shots annotated with the j-th concept's presence as well as shots annotated with its absence, we tune the SVM parameters, C^+ and C^-, which penalize mis-classification of the former and latter types of shots, respectively [12]. The second point is that since the number of shots where the j-th concept is present is much smaller than the number of shots where it is absent, putting the same priority on pl_i^{j1} and pl_i^{j0} leads to a classifier that favors the latter type of shot. Thus, pl_i^{j0} is decreased using ρ.

E^{2}M algorithm: Assume $x_i = (x_i^1, \cdots, x_i^M)$ as the vector representation of the i-th shot where the j-th dimension represents the 'complete' presence (or absence) of the j-th concept with no uncertainty, *i.e.*, $x_i^j \in \{P, A\}$. Clearly, obtaining x_i is impossible because we only have uncertain concept detection scores. The best we can do is to estimate the probability and plausibility of $x_i^j = P$ or $x_i^j = A$. The plausibility is modeled as either pl_i^{j1} or pl_i^{j0}, based on DST. We use the likelihood $L(\theta; pl_i)$ for a given pl_i, which is the set of plausibilities computed for x_i [9]:

$$L(\theta; pl_i) = \sum_{x_i \in \Omega} p(x_i; \theta) pl_i(x_i), \tag{2}$$

where Ω is the domain in which x_i is defined as an M-dimensional vector, and $p(x_i; \theta)$ is the probability that x_i is observed (or generalized) based on the probabilistic distribution with the parameter θ. Equation (2) shows that $p(x_i; \theta)$ represents the imprecision resulting from the population of concept detection scores, while $pl_i(x_i)$ represents the uncertainty related to the error in concept detectors. By assuming that each shot is independently and identically distributed, equation (2) can be extended to N shots:

$$L(\theta; pl) = \prod_{i=1}^{N} L_i(\theta; pl_i) = \prod_{i=1}^{N} \sum_{x_i \in \Omega} p(x_i; \theta) pl_i(x_i), \tag{3}$$

E^{2}M algorithm proposed in [9] computes θ that maximizes $L(\theta; pl)$ given plausibilities for N example shots $(pl_1, \cdots, pl_N)$ based on the Expectation Maximization (EM) algorithm. For reasons of space, we will only describe how E^{2}M algorithm is applied to concept-based video retrieval; please refer to [9] for a complete description. We denote $x_i^{j1} = 1$ if the j-th concept is present in x_i and otherwise $x_i^{j1} = 0$. Similarly, $x_i^{j0} = 1$ if it is absent in x_i, and otherwise $x_i^{j0} = 0$. Assuming that x_i^j is independent from each other, $p(x_i; \theta)$ is written as follows:

$$p(x_i; \theta) = \prod_{j=1}^{M} \prod_{h=0}^{1} (\alpha^{jh})^{x_i^{jh}}, \tag{4}$$

where $\theta = \{\alpha^{jh}\}$ is the set of parameters representing the probability of the j-th concept's presence (α^{j1}) or absence (α^{j0}). By applying Equation (4) to Equation (3), we get:

$$L(\theta; pl) = \prod_{i=1}^{N} \prod_{j=1}^{M} \sum_{h=0}^{1} pl_i^{jh} \alpha^{jh}, \tag{5}$$

where the derivation of $\sum pl_i^{jh} \alpha^{jh}$ is based on the fact that $\sum p(x_i; \theta) pl_i(x_i)$ in Equation (3) can be considered to be the expectation of pl_i [9]. E^{2}M algorithm extracts θ that maximizes the likelihood in Equation (5) as follows:

E-Step: Using θ^q which is the estimate of θ at the q-th iteration, compute $Q(\theta, \theta^q)$, which is the expectation of the log-likelihood of $x = \{x_1, \cdots, x_N\}$:

$$Q(\theta, \theta^q) = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{h=0}^{1} \gamma_i^{jh(q)} \log \alpha^{jh(q)}, \tag{6}$$

 ISBN: 978-1-61208-195-3

where

$$\gamma_i^{jh(q)} = \alpha^{jh(q)} pl_i^{jh} / \sum_{h'=0}^{1} \alpha^{jh'(q)} pl_i^{jh'}. \quad (7)$$

M-Step: Update θ so as to maximize $Q(\theta, \theta^q)$:

$$\alpha^{jh(q+1)} = \sum_{i=1}^{N} \gamma_i^{jh(q)}. \quad (8)$$

Finally, given pl_t of a test shot x_t, we compute its relevance score as the following likelihood $L(\theta; pl_t)$:

$$L(\theta; pl_t) = \prod_{j=1}^{M} \sum_{h=0}^{1} pl_t^{jh} \alpha^{jh}. \quad (9)$$

This likelihood represents the agreement between plausibilities pl_t of x_t and the probabilistic distribution, with θ estimated by E^{2}M algorithm. The set of $1,000$ test shots with the largest $L(\theta; pl_t)$ is returned as a retrieval result.

IV. Preliminary Experimental Results

To test our method, we used TRECVID 2009 video data [5]. This data consists of 219 development and 619 test videos, comprising $36,106$ shots and $97,150$ shots, respectively. We used concept detection scores provided by the City University of Hong Kong, where detection scores for 374 LSCOM concepts are assigned to every shot [6]. Our method was tested for the query "A view of one or more tall buildings and the top story visible". Ten example shots were selected from the development videos. Based on the text description and the example shots, 20 concepts related to the query (*e.g.*, *Building*, *Tower*, *Sky*, *etc.*) were selected using the method in [8].

We conducted a preliminary experiment to examine the effectiveness of using plausibilities, instead of directly using concept detection scores. The following three methods were compared: (1) *DST:* A probabilistic classifier is constructed using plausibilities modeled based on DST, (2) *Sum:* The relevance score of a shot is computed as the sum of concept detection scores (*i.e.*, linear combination with no weights), (3) *Prod:* The relevance score is computed as the product of concept detection scores [2]. Fig. 3 shows a comparison of retrieval performances between *DST*, *Sum* and *Prod* in terms of their average precision. *DST* is superior to the other two. We are now testing *DST* for different queries, and implementing a probabilistic classifier construction method that directly uses concept detection scores.

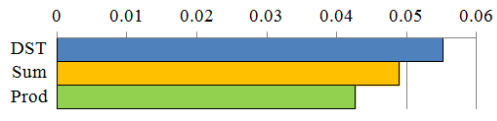

Figure 3. Performance comparison between *DST*, *Sum* and *Prod*.

V. Conclusion and Future Work

In this paper, we introduced a probabilistic classifier construction method which can account for the uncertainty in concept detection by modeling plausibilities of a concept's presence and absence based on DST. We plan to explore the following points to improve the retrieval performance of our method. First, in addition to example shots representing shots that are relevant to a query, we plan to use *counter-example* shots representing irrelevant shots and incorporate them into E^{2}M algorithm. Thereby, irrelevant shots that are retrieved based only on example shots may be excluded from the retrieval result. Second, even for the same query, relevant shots show different combinations of concepts due to varied camera techniques. Thus, we aim to incorporate a mixture model into E^{2}M algorithm. Third, we will explore a method which refines plausibilities of a concept's presence and absence by considering other concepts based on the knowledge base.

References

[1] Natsev A., Haubold A., Tešić, Xie L. and Yan R., "Semantic Concept-based Query Expansion and Re-ranking for Multimedia Retrieval," in *Proc. of ACM MM 2007*, 2007, pp. 991–1000.

[2] Snoek C. et al., "The mediamill TRECVID 2009 semantic video search engine," in *Proc. of TRECVID 2009*, 2009, pp. 226–238.

[3] Ngo C. et al., "VIREO/DVM at TRECVID 2009: High-level feature extraction, automatic video search and content-based copy detection," in *Proc. of TRECVID 2009*, 2009, pp. 415–432.

[4] Naphade M., Smith J., Tešić J., Chang S., Hsu W., Kennedy L., Hauptmann A. and Curtis J., "Large-scale concept ontology for multimedia," *IEEE Multimedia*, vol. 13, no. 3, pp. 86–91, 2006.

[5] Smeaton A., Over P. and Kraaij W., "Evaluation campaigns and TRECVid," in *Proc. of MIR 2006*, 2006, pp. 321–330.

[6] Jiang Y., Ngo C. and Yang J., "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proc. of CIVR 2007*, 2007, pp. 494–501.

[7] Wei X., Jiang Y. and Ngo C., "Concept-driven multi-modality fusion for video search," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 1, pp. 62–73, 2011.

[8] Shirahama K. and Uehara K., "Constructing and utilizing video ontology for accurate and fast retrieval," *International Journal of Multimedia Data Engineering and Management*, vol. 2, no. 4, pp. 59–75, 2011.

[9] Denœux T., "Maximum likelihood estimation from uncertain data in the belief function framework," *IEEE Transactions on Knowledge and Data Engineering*, (PrePrint).

[10] Li X., Wang D., Li J. and Zhang B., "Video search in concept subspace: A text-like paradigm," in *Proc. of CIVR 2007*, 2007, pp. 603–610.

[11] Rasiwasia N., Moreno P. and Vasconcelos N., "Bridging the gap: Query by semantic example," *IEEE Transactions on Multimedia*, vol. 9, no. 5, pp. 923–938, 2007.

[12] Hsu C., Chang C. and Lin C., *A Practical Guide to Support Vector Classification*, http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf retrieved February 2012.

 ISBN: 978-1-61208-195-3