Video Retrieval by Learning Uncertainties in Concept Detection from Imbalanced Annotation Data

Kimiaki Shirahama
College of Information and Systems
Muroran Institute of Technology
shirahama@mmm.muroran-it.ac.jp

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

***Abstract*—Concept-based video retrieval retrieves shots relevant to a query based on detection results of concepts, such as *Person*, *Building* and *Car*. However, concept detection is 'uncertain' because even state-of-the-art methods cannot accurately detect various concepts. Thus, we introduce a video retrieval method, which models the uncertainty in the detection of each concept using 'plausibilities'. A plausibility represents an upper bound of probability that the concept is present (or absent) in a shot. Using such plausibilties, false positive and false negative detections of the concept can be effectively managed. We derive plausibilities by estimating the density ratio between shots annotated with the concept's presence and absence. However, annotating randomly sampled shots does not lead appropriate plausibilities due to the 'imbalanced problem'. This means that the number of shots where the concept is present is generally much smaller than the number of shots where it is absent. To overcome this, a selective sampling method is developed to preferentially sample unannotated shots, which are similar to shots already annotated with the concept's presence. Experimental results on TRECVID 2009 video data validates the effectiveness of derived plausibilities.**

Keywords-Video retrieval; Uncertainty in concept detection; Dempster-Shafer theory; Imbalanced problem; Density ratio;

I. Introduction

Concept-based video retrieval is an approach which retrieves shots relevant to a query based on detection results of concepts, such as *Person*, *Car* and *Building*. Fig. 1 illustrates an overview of concept-based video retrieval. First of all, a *concept detector* is built to detect a concept's presence in shots. Using such detectors, a shot is represented as a multidimensional vector consisting of *concept detection scores*, as shown in Fig. 1 (b). Each detection score represents the probability of a concept's presence. Based on this shot representation, given example shots for a query, a retrieval model is constructed to discriminate between relevant and irrelevant shots to the query. In other words, detection scores for multiple concepts are fused into a single *relevance score*, which indicates the relevance of a shot to the query. Since the detector of a concept is built using a large amount of training shots, the concept can be robustly detected irrespective of its size, position and direction on the screen. Using concept detection scores as 'intermediate' features, concept-based video retrieval can achieve state-of-the-art retrieval performance [1], [2], [3].

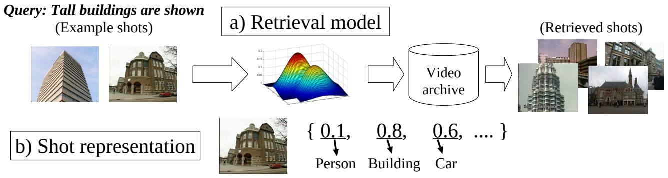

Figure 1. An overview of concept-based video retrieval.

However, even using most effective detectors, it is difficult to accurately detect 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 [1]. At TRECVID 2012, the top-ranked detectors achieved high performances for concepts such as *Male_Person* and *Walking_Running* (with average precisions greater than 0.7). On the other hand, the detection of concepts like *Bicycling* and *Sitting_down* was difficult (with average precisions less than 0.1). Thus, relying on such *uncertain* concept detection significantly degrades retrieval performance.

We have been exploring a method which manages uncertainties in concept detection based on *Dempster–Shafer Theory* (DST) [4]. DST is a generalization of Bayesian theory, where a probability is not assigned to a variable, but instead to a subset of variables [5]. Specifically, we consider two singletons $\{P\}$ and $\{A\}$, which represent the presence and absence of a concept in a shot, respectively. In addition, $\{P, A\}$ represents the uncertainty of whether the concept is present or not. For the above three subsets, a *mass function* m defines masses $m(\{P\})$, $m(\{A\})$ and $m(\{P, A\})$. Here, $m(\{P\})$ and $m(\{A\})$ denote the probability that the concept is certainly present in a shot, and the probability that it is certainly absent, respectively, while $m(\{P, A\})$ denotes the probability that the concept is possibly present in the shot. Using these masses, DST can represent uncertainties much more effective than Bayesian theory, where the only way to represent an uncertainty is to assign the probability 0.5 to both variables P and A.

ISBN: 978-1-61208-265-3

One big difficulty of DST is how to define a mass function. In our case, deriving the mass $m(\{P, A\})$ is substantially infeasible because it is very subjective to annotate shots with $\{P, A\}$ (*i.e.*, a concept's presence is uncertain). Thus, we avoid the mass function derivation by deforming the construction of a retrieval model based on the set-theoretic operation [6]. The retrieval model is constructed based on a *plausibility functions* pl, which is defined by combinations of masses: $pl(\{P\}) = m(\{P\}) + m(\{P, A\})$ and $pl(\{A\}) = m(\{A\}) + m(\{P, A\})$. The plausibility of a concept's presence $pl(\{P\})$ and the one of its absence $pl(\{A\})$ represent upper bound probabilities that it is present and absent in a shot, respectively. Thus, $pl(\{P\})$ is useful for recovering false negative detection of a concept, while $pl(\{A\})$ is useful for alleviating false positive detection.

We mainly address how to derive a plausibility function for each concept. Here, plausibilities of the concept's presence and absence of a shot are obtained based on the detection score of this shot. In our previous work [4], a plausibility function is derived by simple line approximation. However, plausibilities cannot be accurately characterized by lines. Thus, we develop a method which derives a plausibility function by estimating the *density ratio* [7] between shots annotated with a concept's presence and absence on the axis of detection scores. Intuitively, a large plausibility of the concept's presence (absence) should be associated with a detection score, around which the number of shots annotated with its presence (absence) is much larger than that of shots annotated with its absence (presence). Also, plausibilities of the concept's presence and absence should be similar at a detection score, around which numbers of shots annotated with its presence and absence are similar.

However, density estimation involves the *imbalanced problem* [8], meaning that the number of shots where a concept is present is generally much smaller than the number of shots where it is absent. Thus, when annotating randomly selected shots, almost all of them are annotated with the concept's absence, and their detection scores are nearly 0. As a result, an estimated density ratio is much biased towards the detection score 0. To balance numbers of shots annotated with the concept's presence and absence over detection scores, a *selective sampling* method is developed to preferentially select unannotated shots, which are similar to shots already annotated with the concept's presence.

This paper is organized as follows: The next section compares our method to existing ones, in terms of mass and plausibility derivaion, and management in data uncertainty. Section 3 presents our video retrieval method, consisting of retrieval model construction based on DST, plausibility function derivation based on density estimation, and selective sampling. Experimental results in section 4 shows the effectiveness of plausibility functions derived by our method. Section 5 conludes this paper.

II. Related Work

Although several methods for deriving mass and plausibility functions have been proposed, most of them assume special kinds of data like multivariate (transactional) data [9] and data with nested structures [10], or assume an underlying data distribution like Gaussian distribution [11]. Compared to this, we target multi-dimensional categorical data where each dimension represents a concept's presence (P) or absence (A), and does not have any prior knowledge about the data distribution. Hence, we derive plausibility functions in a 'data-driven' approach, where detection scores of shots for the concept are used as source data, and a part of these shots are manually annotated to indicate its presence or absence. In addition, none of existing methods consider the imbalanced problem.

Although an uncertainty in data is addressed in fields of data mining and machine learning, it is defined as a variance of observed values [12]. Compared to this, we define an uncertainty as the inaccuracy of determining the class label of a shot (*i.e.*, a concept's presence or absence). Thus, most of data mining and machine learning methods for uncertain data like [12], cannot be used to deal with uncertainties in this paper.

In concept-based video retrieval, many researchers have explored how to use concept detection scores to achieve accurate retrieval. For example, weighted linear combination is used in [2], [3], where the relevance score of a shot is computed as the sum of weighted detection scores for multiple concepts. Popular weighting methods use the lexical similarity between query terms and a concept, their co-occurrence, and detection scores of the concept in example shots. In [2], a discriminative classifier (*e.g.*, SVM) is built based on the shot representation with concept detection scores. Furthermore, in [13], shots are retrieved based on their similarity to example shots in terms of concept detection scores. To the best of our knowledge, except for our previous work [4], no existing works explicitly address uncertainties in concept detection.

Some researchers addressed uncertainties in combining concept detection results on different features (or modalities) [14], [15]. Such an uncertainty arises when conducting concept detection only using a single feature. In [14], concept detection results on different features are combined based on Portfolio theory, so that for each feature, the expected detection accuracy is maximized and the uncertainty is minimized. Note that this uncertainty is defined as the variance of the detection accuracy on the feature. Compared to this, an uncertainty in this paper means the inaccuracy of detecting a concept's presence or absence. Also, although DST is used in [15], mass function are hand-crafted, so their appropriateness for representing uncertainties is not guaranteed. In this paper, a plausibility function is derived by estimating the density ratio between shots annotated with a concept's presence and absence. This statistically represents the uncertainty of the concept's presence or absence.

III. Video retrieval by modeling uncertainties in concept detection

This section describes our video retrieval method based on DST. First of all, detectors of various concepts are assumed to be already built using a large amount of shots annotated with various concepts' presence and absence. Under this condition, in order to derive a plausibility function for each concept, an additional set of annotated shots are created. In particular, considering the imbalanced problem, our selective sampling method is used to preferentially sample unannotated shots, which are similar to shots already annotated with the concept's presence. Then, a plausibility function is derived by estimating the density ratio between shots annotated with the concept's presence and absence. Finally, given example shots for a query, a retrieval model is constructed by incorporating plausibility functions of different concepts into maximum likelihood estimation.

Below, we first present our video retrieval model where a mass function is transformed into a plausibility function based on DST's set-theoretic operation. Then, our plausibility function derivation and selective sampling methods are described sequentially.

A. Video Retrieval Model based on DST

Our video retrieval model is constructed in the framework of Expectation-Maximization (EM) algorithm [6]. Let $x_i = (x_i^1, \cdots, x_i^M)$ be the 'complete' vector representation of the i-th example shot ($1 \leq i \leq N$). Here, the j-th dimension x_i^j ($1 \leq j \leq M$) represents the presence or absence of the j-th concept with no uncertainty (*i.e.*, $x_i^j \in \{P, A\}$). Assume that x_i^j follows a probability distribution with the parameter θ^j, that is, $p(x_i^j = P; \theta^j)$ and $p(x_i^j = A; \theta^j)$. However, since the detection of the j-th concept is uncertain, $p(x_i^j = P; \theta^j)$ and $p(x_i^j = A; \theta^j)$ incur uncertainties, which are modeled by a mass function m^j. To implement this, based on [6], the following likelihood function $L(\theta; m)$ is used where each example shot and each dimension are assumed to be independent:

$$L(\theta; m) = \prod_{i=1}^{N} \prod_{j=1}^{M} \left(\sum_{S \subseteq \{P,A\}} m^j(S) \sum_{x_i^j \in S} p(x_i^j; \theta^j) \right) \qquad (1)$$

where $\theta = \{\theta^1, \cdots, \theta^M\}$ is a set of parameters for probability distributions for M dimensions (concepts), and $m = \{m^1, \cdots, m^M\}$ is a set of mass functions for M concepts. In addition, S is any subset of $\{P, A\}$, that is, $\{P\}$, $\{A\}$ or $\{P, A\}$. Equation (1) means that $p(x_i^j = P; \theta^j)$ for the complete j-th concept's presence is weighted by masses, which are associated with subsets including P. Similarly, $p(x_i^j = A; \theta^j)$ is weighted by masses, associated with subsets including A. Based on this inclusive relation, the term surrounded by big parenthesis in equation (1) can be expanded and deformed as follows:

$$\begin{aligned} & m^j(\{P\})p(x_i^j = P; \theta^j) + m(\{A\})p(x_i^j = A; \theta^j) \\ & +m(\{P,A\}) \Big(p(x_i^j = P; \theta^j) + p(x_i^j = A; \theta^j) \Big) \\ = \; & p(x_i^j = P; \theta^j) \Big(m^j(\{P\}) + m(\{P,A\}) \Big) \\ & +p(x_i^j = A; \theta^j) \Big(m^j(\{A\}) + m(\{P,A\}) \Big) \\ = \; & p(x_i^j = P; \theta^j) pl^j(\{P\}) + p(x_i^j = A; \theta^j) pl^j(\{A\}) \\ = \; & \sum_{x_i^j \in \{P,A\}} p(x_i^j; \theta^j) pl^j(x_i^j) \qquad (2) \end{aligned}$$

Therefore, the estimation of θ^j does not require the mass function m^j, but requires the plausibility function pl^j. We rewrite $L(\theta; m)$ as $L(\theta; pl)$ where $pl = \{pl^1, \cdots, pl^M\}$ is a set of plausibility functions for M concepts. Estimating θ, which maximizes $L(\theta; pl)$ is equivalent to maximizing the agreement between the probabilistic model $p(x_i^j; \theta^j)$ and uncertain concept detection $pl^j(x_i^j)$.

In our implementation, $p(x_i^j; \theta^j)$ is modeled as a simple discrete probability distribution with two parameters, each of which represents the probability that the j-th concept is present or absent. That is, $\theta^j = \{\alpha^{jP}, \alpha^{jA}\}$. Considering equation (1) and (2), $L(\theta; pl)$ is written as follows:

$$L(\theta; pl) = \prod_{i=1}^{N} \prod_{j=1}^{M} \left(\alpha^{jP} pl^j(x_i^j{=}P) + \alpha^{jA} pl^j(x_i^j{=}A) \right) \qquad (3)$$

Please refer to [4], [6] for the detailed computation process of the estimation of θ. Finally, after θ is obtained using example shots for a query, the relevance score of a test shot x' is computed as follows:

$$rel(x') = \prod_{j=1}^{M} \left(\alpha^{jP} pl^j(x'^j = P) + \alpha^{jA} pl^j(x'^j = A) \right), \qquad (4)$$

where $rel(x')$ represents the agreement between plausibilities of each concept's presence and absence in x' and the probabilistic distribution parameterized by $\theta^j = \{\alpha^{jP}, \alpha^{jA}\}$. The set of $1,000$ test shots with the largest $rel(x')$ is returned as a retrieval result.

B. Plausibility Function Derivation by Density Estimation

For a shot x_i, we compute plausibilities of the j-th concept's presence and absence, $pl^j(x_i^j = P)$ and $pl^j(x_i^j = A)$, based on the detection score of x_i, s_i^j. These plausibilities are defined by the density ratio between two probability distributions, $p_{pr}(s_i^j)$ and $p_{ab}(s_i^j)$. The former represents the probability of the j-th concept's presence at the detection score s_i^j, while the latter represents the probability of its absence at s_i^j.

To compute s_i^j, a concept detector is built as follows: First, each shot is represented using the $1,000$-dimensional Bag-of-Visual-Words representation, where each dimension

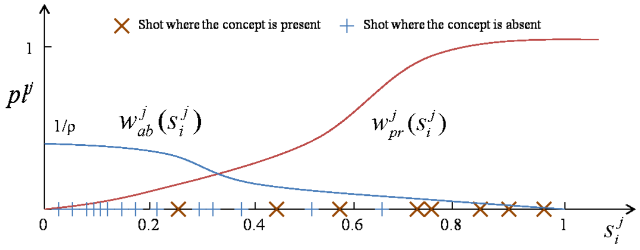

Figure 2. Plausibility computation using density ratio functions

represents the frequency of a characteristic local shape in the keyframe of the shot. Using training shots annotated with the j-th concept's presence and absence, a Support Vector Machine (SVM) is built as a concept detector. The detection score s^j is computed as the SVM's probabilistic output, which approximates the distance between x_i and the detection boundary using a sigmoid function [16].

Fig. 2 illustrates how to compute $pl^j(x_i^j = P)$ and $pl^j(x_i^j = A)$ based on s_i^j. The horizontal axis represents detection scores where ×s represent detection scores of shots annotated with the j-th concept's presence, and +s represent detection scores of shots annotated with its absence. The vertical axis represents plausibilities defined by the following density ratio functions:

$$pl^j(x_i^j = P) = w_{pr}^j(s_i^j) = p_{pr}(s_i^j)/p_{ab}(s_i^j) \quad (5)$$

$$pl^j(x_i^j = A) = w_{ab}^j(s_i^j) = p_{ab}(s_i^j)/p_{pr}(s_i^j) \quad (6)$$

As shown in Fig. 2, $pl^j(x_i^j = P)$ becomes large as a detection score where the number of ×s is larger than the number of +s. On the other hand, $pl^j(x_i^j = A)$ becomes large as a detection score where the number of +s is larger than the number of ×s.

To estimate the density ratio functions $w_{pr}^j(s_i^j)$ and $w_{ab}^j(s_i^j)$, we use the method called *unconstrained Least-Squares Importance Fitting* (uLSIF) [7]. Using uLSIF, $w_{pr}^j(s_i^j)$ is estimated without estimating $p_{pr}(s_i^j)$ or $p_{ab}(s_i^j)$. Instead, it is modeled as the following linear combination of basis functions:

$$w_{pr}^j(s_i^j) = \sum_{l=1}^{b} \alpha_l^j \phi_l(s_i^j), \quad (7)$$

where a weight α_l^j for the l-th basis function $\phi_l(s_i^j)$ is estimated using shots annotated with the j-th concept's presence and absence. We define ϕ_l as a gaussian function. Please refer to [7] for the estimation of α_l^j. Finally, $w_{ab}^j(s_i^j)$ can be obtained in the same way to $w_{pr}^j(s_i^j)$.

C. Sampling from imbalanced data

For appropriate density ratio estimation, we need to solve the imbalanced problem between shots where a concept is present and shots where it is absent. To this end, we present *k-NN based Selective sampling method for Imbalanced Data* (kNNSID). Shots selected by kNNSID are annotated by a user, and used in the density estimation.

Algorithm 1 k-NN based Selective sampling method for Imbalanced Data (kNNSID)

```
Input: $D$: A set of unannotated shots,
       $P$: A set of shots annotated with a concept's presence,
       $N$: Number of shots to be sampled
Output: $R$ : Array that contains sampled shots
1: $D' \leftarrow \mathrm{findUniqueDetectionScores}(D)$
2: while $|R| < N$ do
3:     for all $x \in D'$ do
4:         $score \leftarrow \mathrm{computePriorityScore}(x, R, P)$
5:     end for
6:     $R \leftarrow \mathrm{getTopScoreShot}(D', score)$
7: end while
8: return $R$
```

Figure 3. k-NN based Selective sampling method for Imbalanced Data

Fig. 3 shows a pseudo code of kNNSID, consisting of the following three steps: The first step at line 2 in Algorithm 1 creates a set of unannotated shots, where only one shot is retained for a unique detection score. In the second step at line 7, for each shot, the *priority score* which represents the priority of sampling is calculated. The third step at line 10 samples the shot with the highest priority score. As shown in lines from 4 to 11, the second and third steps are repeated by re-calculating the priority score of each shot until the number of sampled shots reaches the specified number.

The second step calculates the priority score of an annotated shot x, $p(x)$, using the following equation:

$$p(x) = \frac{1}{k_1}\sum_{i=1}^{k_1} d(x, X_i) - \frac{1}{k_2}\sum_{j=1}^{k_2} d(x, Y_j), \quad (8)$$

where $X = \{X_1, X_2, \ldots, X_{k_1}\}$ is a set of already sampled shots that are similar to x. On the other hand, $Y = \{Y_1, Y_2, \ldots, Y_{k_2}\}$ is a set of shots that are similar to x and already annotated with the concept's presence. The function d represents the Euclidean distance between two shots in terms of their detection scores. The first term in equation (8) computes the average distance between x and X. This is useful for collecting shots with a diversity of detection scores. The second term computes the average distance between x and Y. This gives high priorities to shots, which are similar to shots already annotated with the concept's presence. Hence, by annotating sampled shots, we can examine inaccuracies of different detection scores, which are similar to those of shots already annotated with the concept's presence. As a result, we can accurately estimate the density ratio function by alleviating the influence of too many shots where the concept is absent.

IV. Experimental Results

This section evaluates our video retrieval method. First of all, we use 346 concepts defined in Large-Scale Concept Ontology for Multimedia (LSCOM) [17]. These concepts are defined 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 a large mount of training shots. We collect training shots via our online video annotation game [18], which is being developed in parallel with this paper. The game aims to efficiently annotate a large amount of shots with various concepts' presences and absences, with the help of numerous online game users. Specifically, $292,911$ shots in TRECVID 2011 development videos are targeted by the game, and annotated shots are used as training shots to build concept detectors.

The following experiment is conducted by applying the above concept detectors to TRECVID 2009 video data, consisting of $36,106$ shots in 219 development videos, and $97,150$ shots in 619 test videos. For each concept, a plausibility function is derived by the density ratio estimation on $1,000$ shots, annotated with the concept's presence or absence. These shots are collected from development videos using our selective sampling method. Our video retrieval method are tested on the following three queries: (1) "A view of one or more tall buildings and the top story visible", (2) "One or more people, each at a table or desk with a computer visible", and (3) "An airplane or helicopter on the ground, seen from outside". For each query, a retrieval model is constructed using 10 example shots selected from development videos, and used to retrieve relevant shots in test videos. Here, concepts unrelated to the query are ignored to improve the retrieval performance. In other words, concepts related to the query are selected as the ones, for which average detection scores in example shots are larger than the threshold. The retrieval is conducted using detection scores and plausibility functions for selected concepts.

In order to examine the effectiveness of plausibility functions, the above retrieval method denoted by *PL* is compared to a method, which is denoted by *Direct* and constructs a retrieval model directly from concept detection scores. In other words, the model in *Direct* is constructed by replacing $pl^j(x_i^j = P)$ in equation (3) with the detection score s_i^j ($pl^j(x_i^j = A)$ is replaced with $1 - s_i^j$). Fig. 4 shows a performance comparison between *PL* and *Direct* in terms of their precisions. A precision represents the probability of relevant shots in $1,000$ retrieved shots. In each bar graph in Fig. 4, white-colored and black-colored bars represent precisions obtained by *PL* and *Direct*, respectively. In addition, the white-colored and black-colored bars at the top respectively present precisions obtained by plausibility functions (*PL*) and detection scores (*Direct*) for 'ALL' concepts. Each of the other bars presents the precision obtained by the plausibility function (or detection scores) for a single concept. Its name is shown in the left side of the bar graph.

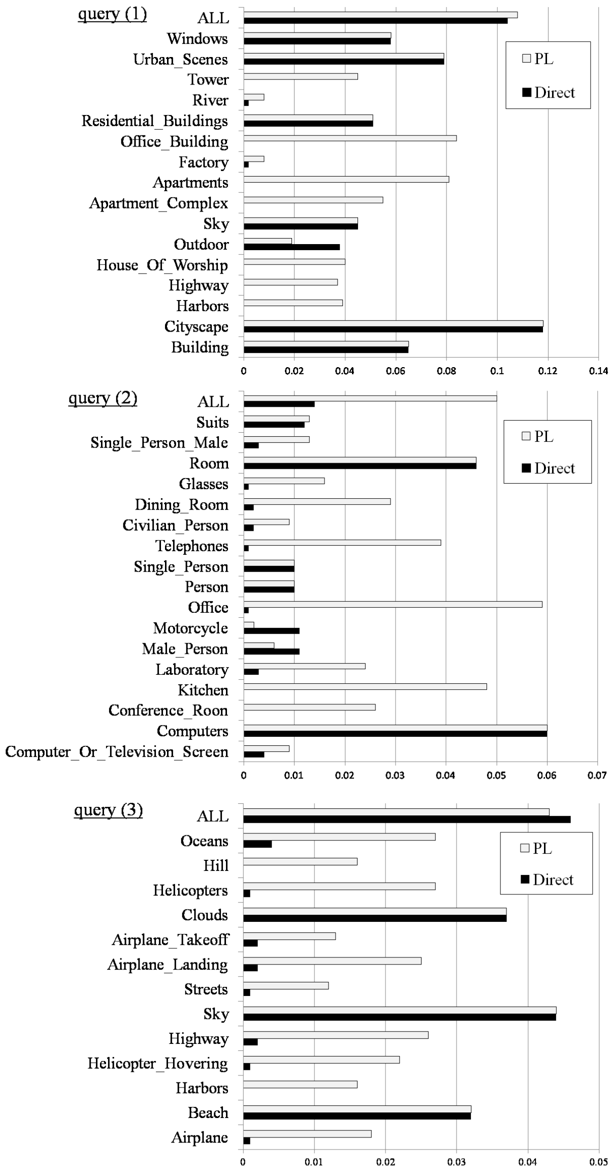

Figure 4. Performance comparison between *PL* and *Direct*

As can be seen from Fig. 4, for query (1) and (2), *PL* is superior to *Direct* in the case of using all concepts. Regarding cases of using single concepts, for almost all concepts where precisions of *Direct* are very low, *PL* achieves much higher precisions. It can be said that detecting such concepts involves much uncertainties, which are effectively modeled by plausibility functions.

However, for query (3), *PL* is outperformed by *Direct* in the case of using all concepts, although precisions of the former are much higher than those of the latter in cases of using single concepts. This means that *PL*'s advantage over *Direct* in cases of using single concepts is weaken in the case of using combinations of these concepts. One main reason is the simplicity of our video retrieval model, where relevant shots to a query are characterized only by a single combination of concepts' presences and absences (see equation (3)). But, actually, relevant shots show different combinations of concepts' presences and absences depending on varied camera techniques. Thus, we plan to incorporate a mixture model into our video retrieval model, or adopt another method, which can extract a non-linear classification boundary between relevant and irrelevant shots based on plausibility functions [19].

V. Conclusion and Future Works

In this paper, we introduced a concept-based video retrieval method where uncertainties in concept detection are modeled using plausibility functions. Each of them is derived by estimating the density ratio between shots annotated with a concept's presence and absence. In particular, to solve the imbalanced problem between the number of shots where the concept is present and that of shots where it is absent, the selective sampling method *kNNSID* is developed to preferentially sample unannotated shots, which are similar to shots already annotated as the concept's presence. Experimental results on TRECVID 2009 video data show that derived plausibility functions effectively manage uncertainties in concept detection. In the future, we plan to improve the retrieval performance in the case of combining plausibility functions for multiple concepts. To this end, our video retrieval method will be extended by incorporating a mixture model, or adopting a method which extracts a non-linear classification boundary between relevant and irrelevant shots to a query using plausibility functions [19].

References

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

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

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

[4] K. Shirahama, K. Kumabuchi, and K. Uehara, "Video retrieval by managing uncertainty in concept detection using dempster-shafer theory," in Proc. of MMEDIA 2012, April, 2012, pp. 71–74.

[5] G. Shafer, A Mathematical Theory of Evidence. Princeton University Press, 1976.

[6] T. Denœux, "Maximum likelihood estimation from uncertain data in the belief function framework," IEEE Transactions on Knoledge and Data Engineering, vol. 25, no. 1, January, 2013, pp. 119–130.

[7] T. Kanamori, S. Hido, and M. Sugiyama, "A least-squares approach to direct importance estimation." Journal of Machine Learning Research, vol. 10, no. 7, July, 2009, pp. 1391–1445.

[8] H. He, and E. Garcia, "Learning from imbalanced data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, September, 2009, pp. 1263–1284.

[9] H. Wang, and S. McClean, "Deriving evidence theoretical functions in multivariate data spaces: A systematic approach," IEEE Transactions on Systems, Man and Cybernetics - Part B, vol. 38, no. 2, March, 2008, pp. 455–465.

[10] A. Aregui, and T. Denœux, "Constructing consonant belief functions from sample data using confidence sets of pignistic probabilities," International Journal of Approximate Reasoning, vol. 49, no. 3, November, 2008, pp. 575–594.

[11] M. Zribi, "Parametric estimation of dempster-shafer belief functions," in Proc. of ISIF 2003, July, 2003, pp. 485–491.

[12] C. Aggarwal, and P. Yu, "A survey of uncertain data algorithms and applications," IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 5, May, 2009, pp. 609–623.

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

[14] X. Wang, and M. Kankanhalli, "Portfolio theory of multimedia fusion," in Proc. of ACM Multimedia 2010, October, 2010, pp. 723–726.

[15] R. Benmokhtar, and B. Huet, "Perplexity-based evidential neural network classifier fusion using MPEG-7 low-level visual features," in Proc. of MIR 2008, October, 2008, pp. 336–341.

[16] C. Chang, and C. Lin, "Libsvm : A library for support vector machines," ACM Transactions on Intelligent Systems and Technology, vol. 2, no. 3, April, 2011, pp. 1–27.

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

[18] Y. Watanabe, K. Shirahama, and K. Uehara, "Online video annotation game with active learning and tag ranking," IEICE Technical Report, vol. 112, no. 346, December, 2012, pp. 75–80 (In Japanese).

[19] T. Denœux, "A k-nearest neighbor classification rule based on dempster-shafer theory," IEEE Transactions on Systems, Man, and Cybernetics, vol. 25, no. 5, May, 1995, pp. 804–813.