A Multimedia Retrieval Framework Based on Automatic Graded Relevance Judgments

Miriam Redi and Bernard Merialdo

Eurecom, Sophia Antipolis
{redi,merialdo}@eurecom.fr

Abstract. Traditional Content Based Multimedia Retrieval (CBMR) systems measure the relevance of visual samples using a binary scale (Relevant/Non Relevant). However, a picture can be relevant to a semantic category with different degrees, depending on the way such concept is represented in the image. In this paper, we build a CBMR framework that supports graded relevance judgments. In order to quickly build graded ground truths, we propose a measure to reassess binary-labeled databases without involving manual effort: we automatically assign a reliable relevance degree (Non, Weakly, Average, Very Relevant) to each sample, based on its position with respect to the hyperplane drawn by support vector machines in the feature space. We test the effectiveness of our system on two large-scale databases, and we show that our approach outperforms the traditional binary relevance-based frameworks in both scene recognition and video retrieval.

1 Introduction

CBMR systems aim to categorize and search for visual content in large collection of visual data, by exploiting statistical models that predict the presence of semantic concepts in images or videos. Such frameworks generally rely on supervised learning techniques, that require manually-assessed ground truth annotations associated with the samples in the dataset. When labeling a dataset, real assessors are asked to categorize an image or a shot according to its topical relevance with respect to a given concept. In most cases (e.g. the TrecVid collaborative annotation [1]), the notion of relevance is measured using a binary scale: a visual input is either "positive" or "negative" for the concept considered. This type of assessment assumes therefore that all the relevant elements are identically relevant and that all the irrelevant samples are equally non-relevant.

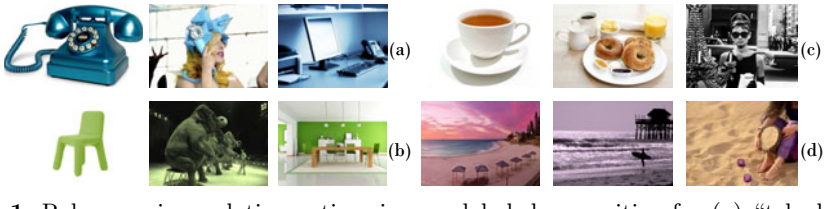

Fig. 1. Relevance is a relative notion: images labeled as positive for (a) "telephone" (b) "chair" (c) "cup" (d) "beach" actually have different visual evidences

K. Schoeffmann et al. (Eds.): MMM 2012, LNCS 7131, pp. 300–311, 2012.

However, "as all human notions, relevance is messy and not necessarily perfectly rational" [16]. Each group in Fig. 1 shows a set of images that would be annotated as positive for the same corresponding concept: even if we can acknowledge that all the images are relevant with respect to the group label (e.g. images in group *a* contain the concept "Telephones"), the global semantic content of each image differs. Intuitively, we would say that each image is relevant for the associated concept with a different degree (for example, similar to web search engines, labels or grades such as "weakly relevant" or "very relevant" could be assigned). A distribution of relevance inferences over a graded scale would reflect better the human way of understanding concepts. From a learning system point of view, binary judgments imply that both marginally-relevant samples and very representative samples are treated equally when modeling the concept feature space: this might cause inconsistencies in the classification process. In a multimedia retrieval framework, concept models might be therefore less effective due to the contrast between the intra-class diversity and the binary relevance judgment. While graded relevance is widely used in web information retrieval (see [4], [23]), its use was rarely explored in CBMR: an attempt is represented by the graded-relevance system of Elleuch et Al. [3], that in the TrecVid 2010 edition outperformed the traditional binary-relevance frameworks proposed by the other participants at the Semantic Indexing Task.

When building a graded-relevance framework for information retrieval, the first step is to reassess the training samples, labeled as positive/negative, by assigning a "degree" of relevance. Generally [19] [3], the level of relevance of each sample is labeled manually. However, when dealing with large collections of visual data, e.g. the 400 hours of training videos for TrecVid [17] 2011, such re-assessment becomes time-consuming and practically unfeasible.

In this paper we propose an effective automatic graded-relevance based framework for image recognition and video retrieval. With our system, we can treat noisy and marginally relevant samples with less importance, achieving a better usage of our training set, thus improving the performances of traditional binary-relevance systems. Moreover, the key aspect of our framework is that, unlikely [3], the relevance degree of a training sample is assessed automatically: we assign to each sample a reliable and realistic relevance judgment, without involving any manual effort. To auto-reannotate each training sample in the database according to a non-binary relevance scale, we find a measure that first assigns a fuzzy membership judgment (i.e. how much a sample is representative/positive for a given concept), based on the position of the sample with respect to the hyperplane drawn by a Support Vector Machine (SVM) [2] in the feature space. Then, based on such relevance score, we re-categorize the training dataset into 4 groups for every concept: Very Relevant, Average Relevant, Weakly Relevant and Non Relevant samples. By training the system on such multiple repartitions, we then build a multi-level model for each semantic concept considered. When assigning labels to a new sample, the system outputs a set of concept prediction scores (one for each relevance-based layer of the model), that we weigh and combine to obtain a final label.

We test the effectiveness of our system by comparing it with traditional binary-relevance frameworks in two different tasks, namely scene categorization and video retrieval. For the first task we consider a large scale, noisy, database of tourism-related images, and we show that traditional categorization systems and features benefit from our automatic graded relevance-based multi-layer model when classifying this kind of biased data. We also consider the non-trivial Semantic Indexing Task of TrecVid 2010 [17] and we show that our non-binary reassessment combined with a multi-level prediction improves the recognition performances of a binary-scale video retrieval system by about 13%.

The remainder of this paper is structured as follows: in Sec. 2 we present an overview of the related work; in Sec. 3 we outline some background knowledge on traditional SVM-based retrieval systems; in Sec. 4 we show how to build an automatic relevance degree assignment scheme in a video retrieval framework. Finally, in Sec. 5 we compare our proposed framework with traditional image recognition and video retrieval systems and evaluate the results.

2 Related Work

Relevance is a fundamental notion for information retrieval: as pointed out in [16], while traditional bibliographic and classification frameworks aim to describe/categorize samples, retrieving information involves, besides description and categorization, the need for *searching*, and "searching is about relevance". Graded relevance-based learning methods first appeared for real Web search engines, where pages cannot be simply categorized as relative/non relative, but they need a multi-level relevance assignment. Several algorithms have been proposed to learn ranking functions from relative relevance judgments, like RankNet[21], based on neural networks, RankBoost[4], or the regression-based learning proposed in [23] by Zheng et al. How are these "grades" assigned? Generally, in traditional information retrieval such reassessment is done manually, either using real expert assessors [19], or using Amazon MechanicalTurk [18]. For web-based searches, the relevance judgment can be inferred in an automatic way, using the users' clickthroughs (see [7] for an overview of implicit relevance feedback method). In the image analysis and video retrieval field, graded relevance has been rarely explored. Traditional multimedia retrieval systems (see, for example, [14]) generally rely on binary-labeled keyframes or images. However, it was recently shown [3] that a video retrieval framework benefits from a graded-relevance annotated training set: in [3] the development set is reassessed by assigning, for each generally "relevant" frame, a degree of relevance from Somehow Relevant to Highly relevant. Three new training sets are then created based on different combinations of the relevance-based partitions.

Our work is somehow similar to the framework presented in [3]; however, in their work, the manual database re-assessment involves a lot of human effort and might increase the labeling noise. In this paper we automatize this process by automatically assigning a class membership degree to each sample. The idea is to exploit the learning methods traditionally used in video retrieval frameworks: the SVMs. Few works have indeed been presented in machine learning

literature that reassess the samples in a binary-labeled training set based on the learnt feature space. Generally, they assign to the samples automatically a fuzzy membership score, namely a value representing their relevance for a given class. For example [9] defines an automatic membership measure as a function of the mean and radius of each class; this work is then extended by Lin et al in [8], that uses an heuristic strategy based on a confidence factor and a trashy factor in training data to automatically assign a score to each sample. An example of using automatic relevance assignment for image recognition is represented by the work of Ji et al. [5], where, to solve a face gender classification problem, the distance to the SVM hyperplane is used to measure the importance of each sample in a dataset for a given class. Another example can be found in [12], where the confidence of an image region label is again derived from the sample distance from the hyperplane. Similar to the work in [5], we use a SVM-based measure to identify a fuzzy relevance score for each class, that we then discretize, in order to label our training sets with three relevance degrees. However, instead of using the raw distance value, we prefer to use a calibrated, thresholded value, that still depends on the distance to the hyperplane, but it is expressed with the probability of a given sample to be positive with respect to a concept.

3 Binary Relevance Based Retrieval Systems

Traditional multimedia categorization systems associate a set of images or videos with a semantic label given a low-dimensional description of the input, namely a feature vector. Multimedia retrieval systems use categorization frameworks to build lists of pictures/shots ranked according to their pertinence with respect to a semantic concept or query. In both cases, the problem can be reduced to a multiclass classification problem, where each class represents the query/concept to be found in a visual sample. Generally, concept-specific SVMs are used to build models able to predict the presence of a given concept in a visual sample.

In order to build such system, a set of training samples $(x_i, y_i l)$, $i = 1 \ldots n$ is required, where $x_1, x_2, \ldots, x_n$ are the feature vectors extracted from the visual input data, and $y_i l$ the associated labels. For a set of concepts or categories $\{c_1, c_2, \ldots . c_p\}$ (e.g. "Telephones"), each sample is labeled either as "positive", $y_{il} = +1$, $l = 1, \ldots, p$, (the concept is present in the visual input represented by x_i) or "negative", $y_{il} = -1$ (no visual trace of the concept is found in x_i). A set of SVM-based classifiers, one for each concept/category, is used to learn the feature space and then to label new samples according to the same scheme. The idea behind the SVM is to find a hyperplane that optimally separates the two classes ($y_{il} = \pm 1$) in the problem feature space, given the distribution of the positive and negative samples with respect to a given concept. Such hyperplane satisfies the equation $\sum_i (\alpha_{il} y_{il} x_i)^T x - b_l = 0$.[1] When a new sample z needs to be categorized, the system assigns the corresponding label y_{zl} based on the

[1] Where $w_l = \sum_i \alpha_{il} y_{il} x_i$ has been proved in [2] to be the linear combination of the support vectors (i.e. the samples x_i for which the corresponding Lagrangian multiplier α_i is non zero).

sign of the dot product-based decision function $f_l(z) = w_l^T z + b_l$. For a retrieval framework (see, for example [14]), a confidence score $p(y_{zl} = 1|z)$ is obtained for sample z based on decision function values. Generally, a set of v visual features are extracted from each sample. v scores are obtained given such features, and then combined into one single confidence score. The results are then ranked according to such final score.

4 A Graded Relevance Based Retrieval System

As showed in Sec. 3, a SVM separates the feature space so that we are able to distinguish between positive and negative new samples for each given concept. This boundary is found based on a binary relevance judgment, y_{il}, that, as discussed before, might be too restrictive compared to the range of possible instances of a semantic concept in the visual input. In order to allow a better usage of our data, we go beyond the Relevant/Irrelevant subdivision, by reassessing our binary-relevance based training set with graded relevance judgments: in the new training set, a frame can be either Irrelevant (negative), Weak/Marginally Relevant, Average Relevant or Very Relevant. We then integrate the inferred relevance degree in a multi-layer concept classifier. The proposed framework works as follows (see Fig. 2):
(1)The features extracted from the training samples are processed by a set of binary p SVM-based classifiers (one for each concept). According to such models, we analyze the position of each training sample x_k with respect to the hyperplane, using a calibrated decision value, and extract, for each concept c_l, a fuzzy membership score σ_{kl}. This is a continuous value representing how much a given sample is representative for a semantic concept (see Sec 4.1 for more details).
(2) As shown in Section 4.2, for each concept, we sort the positive training samples according to their fuzzy relevance scores and we set two thresholds so that we are able to re-categorize the samples using discrete relevance degrees. We obtain three subsets of Strongly, Average and Weakly Relevant training samples. All the negatives are equally labeled as Non Relevant samples.
(3) Similar to [3], we then build a multi-layer model by training the system on three different, relevance-based training sets. Then, as presented in Sec 4.3, given a new test image, for all c_l we obtain from the multi-layer model three different concept prediction scores, that we then combine with weighted linear fusion to obtain one single output score. Such output score is then used for ranking and thresholded to determine the image label.

4.1 Decision Values as Relevance Indicators

As any traditional retrieval system, we start from an annotated training set of images/keyframes represented using low level features, namely our labeled samples. Given a set of non-negative samples, how to automatically define the fuzzy degree of relevance σ_{kl} of each sample with respect to a semantic concept? We tackle this problem by exploiting the SVM decision values of the training

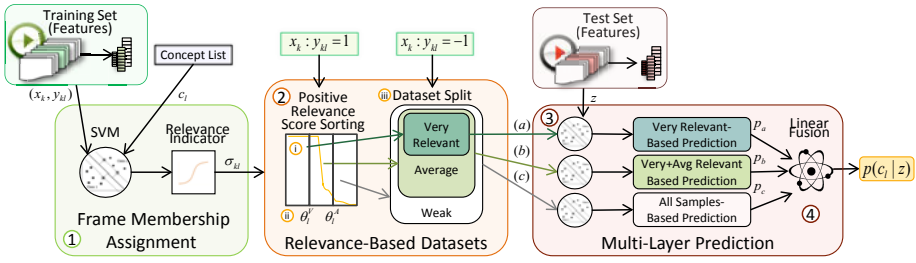

Fig. 2. Visual representation of our Relevance Based Framework

set. The idea is that if, for a concept c_l, we are able to define how "positive" the sample is, given its position with respect to the hyperplane, we can have a good estimation of its relevance degree for that given concept. As a matter of fact, various works [5,9,12] showed that there is a correlation between the distance to the hyperplane (or the distance to the class center) and how much each sample is representative for a given class (the bigger its distance from the boundary, the higher its relevance with respect to the positive/negative category).

In our approach, we use as a fuzzy membership measure for a training sample a thresholded version of the decision function, according to the solution proposed in [13] to translate the uncalibrated decision value into a probabilistic output. First, we calculate $f_l(x_k)$, namely the decision value for concept c_l, $\forall x_k, k = 1, \ldots, n$ in the training set samples. We then estimate the membership assignment as the positive class posterior probability $\sigma_{kl} = p(y_{kl} = 1 | f_l(x_k))$ with a parametric model based on fitting a sigmoid function:

$$\sigma_{kl} = \frac{1}{1 + \exp(A f_l(x_k) + B)}, \tag{1}$$

Where A and B are parameters adapted in the training phase to give the best probability estimates.

4.2 A Multi-layer Training Set with Different Relevance Levels

Once the *continuous* value σ_{kl} is computed for each training sample x_k, the next step is to build a graded relevance retrieval framework. In order to achieve this goal, we need to have a *discrete* relevance degree for each training sample, so that we are able to perform a relevance-based split of the training set into smaller, consistent subsets with different degrees of relevance with respect to a concept c_l. As pointed out in [6], there is no universal rule to define such number of relevance degrees in a graded system. However, as shown in Sec 5.2, our experimental results suggest to set to 4 the number of relevance levels considered.

We therefore separate, for each concept, the positive/relevant training samples into three groups: Very Relevant Samples, that represent the most representative images/keyframes for a given class, Average Relevant Samples, and Weakly Relevant Samples; all the negatives are equally labeled as Non Relevant samples.

We then generate three repartitions of our training database, based on which a multi-layer model will be learnt (see Sec. 4.3). Having the fuzzy membership score σ_{kl} for each relevant sample, the discretization procedure is very simple:

(i) For each c_l, we take the *positive* ($x_k : y_{kl} = 1$) training samples and sort them according to their corresponding σ_{kl}, in decreasing order.
(ii) We now want to find a partition of the positive samples in three classes, according to the relevance scale selected. Based on the shape of the curve drawn by the sorted fuzzy relevance scores, we identify two thresholds, θ_l^V and θ_l^A. We use and test three different approaches to choose such thresholds: (ii.a) we split the curve into equally spaced intervals, (ii.b) we choose the thresholds manually such that, intuitively, the intra-partition variance of the scores value is minimized (ii.c) we choose the values corresponding to 1/3 and 2/3 of the maximum membership score for the concept considered . For each concept c_l, the Very Relevant samples are then defined as the positive $x_k : 1 < \sigma_{kl} < \theta_l^V | y_{kl} = 1$; the Average Relevant samples as $x_k : \theta_l^V < \sigma_{kl} < \theta_l^A | y_{kl} = 1$; the Weakly Relevant as $x_k : \theta_l^A < \sigma_{kl} < 0 | y_{kl} = 1$.
(iii) Finally, similar to [3] we create three new training sets: (a) merges the Very Relevant Samples with all the Non Relevant (i.e. our *negatives*, $x_k : y_{kl} = -1$), (b) merges (a) with the Average Relevant Samples, and (c) considers all positives and negatives samples.

4.3 Multi-layer Prediction and Fusion

Once we have created the three concept-specific training subsets, for each concept we build our multi-layer model: it consists of three different SVM-based models, each of them learning a partition (a), (b), (c). Each level of the model separates the feature space in a different way, according to the annotations of the subset considered. When a new test sample z needs to be classified, we compute, using probabilistic SVM, three prediction scores for each concept (each of them is generated by a layer of the model). We therefore obtain , $\forall l$, $p_a(y_{zl} = 1|z), p_b(y_{zl} = 1|z), p_c(y_{zl} = 1|z)$.

Each of these predictions is generated by a different relevance-based partition, which gives a different, complementary type of information regarding the relevance degree of the new sample to be classified. In order to exploit such different cues and obtain a single output, we then merge the three outputs using weighted linear fusion, as follows:

$$p_{zl} = p(y_{zl} = 1|z) = \sum_t w_t p_t(y_{zl} = 1|z), \tag{2}$$

$t = a, b, c, \forall l$, where w_t is a concept-specific weight learnt with development data. For retrieval purposes, we then rank, for each query l the test samples according to p_{zl} in decreasing score, while for image categorization, the final label y_{zl} is assigned according to the following scheme:

$$y_{zl} = \begin{cases} -1 & \text{if } p_{zl} < 0.5 \\ +1 & otherwise \end{cases}$$

5 Experimental Validation

In this section, we use our proposed framework for both scene recognition and video retrieval: we compare the graded relevance framework with the classical binary-relevance systems (our baselines) for both tasks. First, in Sec. 5.1 we briefly summarize the composition of the large-scale databases considered and the experimental setup of the binary-relevance systems we use as baselines. We then explain in Sec 5.2 some details about our graded relevance framework setup and present some visual results that validate our automatic membership measure, presented in Sec 4.1. Finally, in Sec 5.3 we present the results obtained by comparing binary and graded relevance systems, for both the considered tasks.

5.1 Binary Relevance Framework Setup: Databases and Baselines

Scene Recognition. For this task, an automatic annotation system is required to assign a semantic category/concept to each image in the database. We choose for this task a large-scale database composed of around of 100,000 images coming from 100,000 touristic properties[2]. The database spans 16 between outdoor and indoor scene categories. For our binary-relevance baseline, we extract from such database the most widely used global features for content based image retrieval, namely Color Moments [20] , Wavelet Feature [11], Edge Histogram [22] and Saliency Moments [15] (respectively "CM", "WF", Edge" and "Saliency" in Fig. 3(d)). For every considered feature, a one-versus-all polynomial SVM-based model is built to separate each class from the others. Finally, the label confidence score of all the features are combined with linear fusion to obtain one single output.("all" in Fig. 3(d)).

Video Retrieval. Here, we focus on the Light Semantic Indexing Task (SIN), of TrecVid [17] 2010 where the retrieval system is required to produce a ranked list of relevant shots for a set of semantic concepts proposed. We use as a database the TrecVid 2010 IACC.1.tv10.dev set, which is composed of 3200 Internet Archive videos (a total of around 100,000 shots), that have been annotated with binary assignments. For our baseline, similar to our system in TrecVid 2010 [14], from each keyframe/shot, we extract a pool of visual features (Sift [10], Color Moments [20], a Wavelet Feature [11], and the MPEG7 edge histogram [22]). We then use them as input for a set of concept-specific classifiers, to build models that will predict the presence of a concept in each keyframe, and output a label and a concept score (the label confidence). All the concept scores coming from the different features are linearly combined to obtain the final concept score for each shot, that we will use to build the ranked list of shots.

[2] This is a randomly sampled subsed of 1 million images describing hotels amenities and surroundings, that have been manually labeled on the property owner's side before uploading them into a Hotel Management Platform.

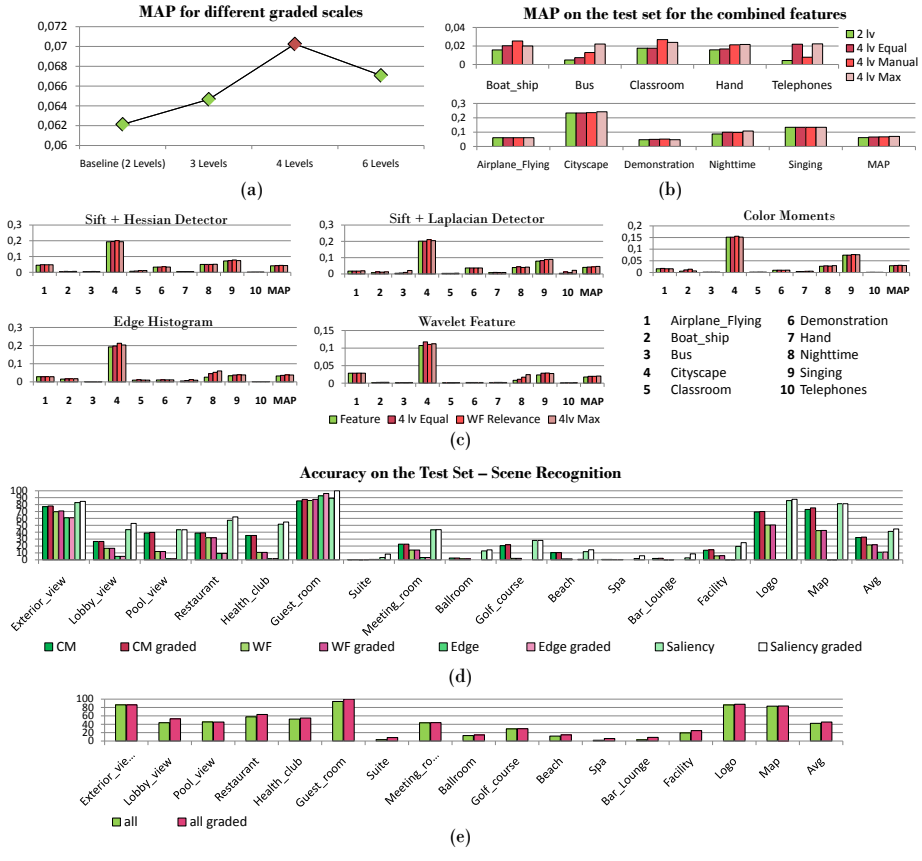

Fig. 3. We compare our system with a traditional binary-relevance CBMR system. Video Retrieval task (a) Mean Average Precision values with different numbers of relevance-based categories (b) per-concept Average Precision on the TrecVid Database given the complete set of features (c) per-feature results. Image Categorization task (d) per-feature results and (e) Average Precision Accuracy on the test set for the combined set of features

5.2 Graded Relevance Framework Setup: Scale Selection and Relevance Visual Results

Our Graded Relevance frameworks are built on top of the baselines outlined in the previous section. As we already have binary annotated datasets, we need to (1) add a fuzzy membership score to each frame, (2) find proper thresholds to obtain a discrete relevance category assignment, and (3) build a multi-layer model as described in Sec 4.1-4.2-4.3.

(1)For each feature f, we can re-use the model built in the baseline to estimate the fuzzy membership score σ_{kl}^{f} of a keyframe/image in the training set x_k for a concept/category c_l. Instead of using directly feature-based membership scores, that might supply incomplete information (e.g. the most relevant samples given

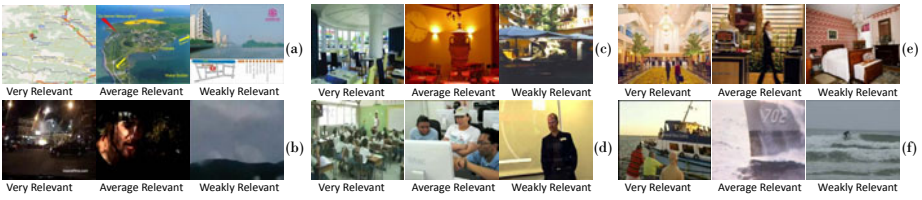

Fig. 4. Automatic relevance-based reassessment: for given semantic concepts, examples from the three relevance-based categories are shown (a)Map (b)Nighttime (c)Restaurant (d)Classroom (e)Hotel Lobby (f)Boat

the color or the edge distribution only), we combine them to obtain one single σ_{kl} for each sample.

(2)Now that we have a fuzzy score, how to select the number of discrete levels that we will use to re-categorize the training set? As shown in Fig. 3, we experimented with different subdivisions of the relevant samples of the training set and tested their respective performances on the video retrieval task. Results shown in Mean Average Precision (MAP) yield to the selection of a 4-level graded scale (namely Highly, Average and Weakly Relevant, and the Non Relevant label assigned to all the negatives) to reassess the training set. Is this subdivision reliable? Fig. 4 shows examples from the three relevance-based classes: as we can see, our proposed method actually separates samples according to their relevance with respect to the given category or query, and in some cases, among the "Weakly Relevant" samples we can even find wrongly annotated images. Given the trend of the fuzzy membership score curve, we select the thresholds θ^V and θ^A, according to methods (ii.a), (ii.b), (ii.c) mentioned in Sec. 4.2 (respectively "4lv Equal", "4lv Manual", "4 lv Max" in Fig. 3 b and c).

(3) Finally, for every feature and every concept, given the new training set repartitions, three models are created and then used to predict the presence of the concept, combining the three outputs as shown in Sec. 4.3. At the end of this step we will have, for a new sample z, a concept score p^f_{zl} for each feature. Such feature-specific concept scores are then fused with linear fusion, similar to the binary baseline.

5.3 Results

Scene Recognition. The scene recognition results in Fig. 3 (d-e) show the improvement obtained on a traditional binary relevance categorization system by introducing our graded-relevance reassessment, evaluated with the standard average classification accuracy on the test set. If we consider the whole set of descriptors combined together ("all" vs "all graded"), we can see that with our system we improve the overall categorization performances of about 8%. In particular, we can see that, when switching to graded relevance, we improve the discriminative power for some particular categories (e.g. Spa, +214%, Bar/Lounge, +197% and Beach, +24%) : analyzing such categories, we saw that those are the classes that are more affected by labeling noise, because they are often confused by the manual assessor with semantically similar classes (e.g. Bar-Restaurant, Spa-Health Club, Beach-Exterior View).

Video Retrieval. For the Video Retrieval Task, we present the results of both systems in terms of Mean Average Precision, the standard evaluation measure used for TrecVid assessments. We can see from Fig. 3 that the weaker features (e.g. Edge Histogram, +20% and Wavelet Feature, + 15%) benefit from our graded system. Moreover, we can see that the overall MAP increases of about 13%, when considering the ensemble of features combined together, with some peaks for those concepts for which the binary system was less performing, e.g. Classroom +53%, Telephones +420%, Bus +356% and BoatShip +60%.

6 Conclusions and Future Work

We presented a Multimedia Categorization and Retrieval Framework based on automatic graded relevance annotations. We automatically reassessed binary-labeled databases by assigning a degree of relevance to each sample based on its position with respect to the SVM hyperplane, and build an effective graded-relevance based CBMR system. We showed that our system, by allowing different degrees of relevance, outperforms the traditional binary-based frameworks for both image recognition and video retrieval.

Our simple approach can be improved in various ways. First, the automatic relevance fuzzy score assignment can be refined by using more complex machine learning-based measures, or by considering the combination of the relevance scores of a sample with respect to different concepts. Moreover, we can automatize the discretization procedure (from fuzzy to discrete relevance degrees) by designing a measure that infers the best thresholds from the shape of the positive membership scores curve. Finally, while in our framework, similar to traditional CBMR systems, we use simple SVM classifiers for ranking, we could explore the learning methods used for web page ranking (e.g. [23]), that are designed to support graded-relevance, achieving a higher discriminative power.

References

1. Ayache, S., Quénot, G.: Trecvid 2007 collaborative annotation using active learning. In: Proceedings of the TRECVID 2007 Workshop (2007)
2. Boser, B.E., Guyon, I.M., Vapnik, V.N.: A training algorithm for optimal margin classifiers. In: Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 144–152. ACM (1992)
3. Elleuch, N., Zarka, M., Feki, I., Ammar, A.B., Alimi, A.: Regimvid at trecvid 2010: Semantic indexing. In: Proceedings of the TRECVID 2010 Workshop (2010)
4. Freund, Y., Iyer, R., Schapire, R.E., Singer, Y.: An efficient boosting algorithm for combining preferences. J. Mach. Learn. Res. 4, 933–969 (2003)
5. Ji, Z., Lu, B.-L.: Gender Classification Based on Support Vector Machine with Automatic Confidence. In: Leung, C.S., Lee, M., Chan, J.H. (eds.) ICONIP 2009, Part I. LNCS, vol. 5863, pp. 685–692. Springer, Heidelberg (2009)
6. Kekäläinen, J.: Binary and graded relevance in ir evaluations–comparison of the effects on ranking of ir systems. Information processing & management 41(5), 1019–1033 (2005)

7. Kelly, D., Teevan, J.: Implicit feedback for inferring user preference: a bibliography. SIGIR Forum 37, 18–28 (2003)
8. Lin, C., Wang, S.: Training algorithms for fuzzy support vector machines with noisy data. Pattern Recognition Letters 25(14), 1647–1656 (2004)
9. Lin, C.F., Wang, S.D.: Fuzzy support vector machines. IEEE Transactions on Neural Networks 13(2), 464–471 (2002)
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
11. Papageorgiou, C.P., Oren, M., Poggio, T.: A General Framework for Object Detection. In: Proceedings of the Sixth International Conference on Computer Vision, p. 555. IEEE Computer Society (1998)
12. Paterno, M.C.S., Lim, F.S., Leow, W.K.: Fuzzy semantic labeling for image retrieval. In: 2004 IEEE International Conference on Multimedia and Expo, ICME 2004, vol. 2, pp. 767–770. IEEE (2004)
13. Platt, J.: Probabilistic outputs for support vector machines. In: Bartlett, P., Schoelkopf, B., Schurmans, D., Smola, A.J. (eds.) Advances in Large Margin Classifiers, pp. 61–74
14. Redi, M., Merialdo, B., Wang, F.: Eurecom and ecnu at trecvid 2010: The semantic indexing task. In: Proceedings of the TRECVID 2010 Workshop (2010)
15. Redi, M., Merialdo, B.: Saliency moments for image categorization. In: Proceedings of the 1st ACM International Conference on Multimedia Retrieval, ICMR 2011, pp. 39:1–39:8. ACM, New York (2011)
16. Saracevic, T.: Relevance: A review of the literature and a framework for thinking on the notion in information science. part iii: Behavior and effects of relevance. Journal of the American Society for Information Science and Technology 58(13), 2126–2144 (2007)
17. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: MIR 2006: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pp. 321–330. ACM Press, New York (2006)
18. Snow, R., O'Connor, B., Jurafsky, D., Ng, A.Y.: Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 254–263. Association for Computational Linguistics (2008)
19. Sormunen, E.: Liberal relevance criteria of trec-: counting on negligible documents? In: Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 324–330. ACM (2002)
20. Stricker, M.A., Orengo, M.: Similarity of color images. In: Proceedings of SPIE, vol. 2420, p. 381 (1995)
21. Svore, K., Vanderwende, L., Burges, C.: Enhancing single-document summarization by combining ranknet and third-party sources. In: Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pp. 448–457 (2007)
22. Won, C.S., Park, D.K., Park, S.J.: Efficient use of MPEG-7 edge histogram descriptor. Etri Journal 24(1), 23–30 (2002)
23. Zheng, Z., Chen, K., Sun, G., Zha, H.: A regression framework for learning ranking functions using relative relevance judgments. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 287–294. ACM (2007)