UNIFYING LOCAL AND GLOBAL CONTENT-BASED SIMILARITIES FOR HOME PHOTO RETRIEVAL

Joo-Hwee Lim

Institute for Infocomm Research
21 Heng Mui Keng Terrace, Singapore 119613
joohwee@i2r.a-star.edu.sg

Jesse S. Jin

University of New South Wales
Sydney 2052, Australia
jesse@cse.unsw.edu.au

ABSTRACT

Unlike professional or domain-specific images, home photos vary significantly. They pose great challenge for content-based image retrieval. In this paper, we propose a Bayesian formulation to unify both local and global content-based similarities for image matching and demonstrate its superior retrieval performance on 2400 genuine home photos. Our proposed framework uses support vector machines to extract and combine intra-image and inter-class semantics. Support vector detectors are first trained on semantically meaningful regions and used to form detection-based image indexes. The indexes then serve as input for support vector learning of image classifiers to generate class-relative indexes. During image retrieval, similarities based on both detection-based and class-relative indexes are combined to rank images. Query-by-example experiments on 2400 home photos with 16 semantic queries show that the combined matching approach is better than matching with single index. It also outperformed the fusion of color and texture features by 55% in average precision.

1. INTRODUCTION

Unconstrained home photos pose great challenge for content-based image retrieval research as the amount of content variations is usually very high due to the spontaneous and casual nature during image capturing. More often than not, the objects in the photos are ill-posed, occluded, and cluttered with poor lighting, focus, and exposure. In particular, a challenge for computer vision is the usually very large number of object classes in polysemic images. Hence detecting semantic objects (e.g. faces, sky, foliage, buildings etc) based on trained pattern classifiers has received serious attention lately (e.g. [2]). Highly accurate segmentation of objects is a major bottleneck except for selected narrow domains when few dominant objects are recorded against a clear background ([4],p.1360). The challenge of object segmentation is acute for polysemic images in broad domains like home photos. A key innovation in our method is that no region segmentation is needed. Instead, visual concepts are learned and detected during image indexing from tessellated image blocks, as inspired by multi-scale view-based object recognition framework [3]. Moreover, local detection decisions are reconciled across multiple resolutions and aggregated over spatial areas as semantic indexes.

On the other hand, image categorization is a powerful divide-and-conquer metaphor to organize and access images. In general, classifications were based on low-level features such as color, edge directions etc and [5] presented a comprehensive coverage of the problem by dealing with a hierarchy of 8 categories (plus 3 "others") progressively with separately designed features. The vacation photos used in their experiments are a mixture of Corel photos, personal photos, video key frames, and photos from the web. In our case, image classification is not the end but a means to compute categorical similarity so as to provide contextual estimation of the relevance class of a query.

While Naphade [2] proposed a probabilistic framework that enhances the detection of probabilistic multimedia objects called multijects in videos by modeling their interrelationship in an explicit network form (multinet), our approach is simpler as both the local semantics and their implicit co-occurrence context are trained separately, hence simplifying the learning problem. In addition, segmented objects and sites (e.g. outdoor scene) are treated as equal entities as multijects [2]. In our case, segmentation-free block regions and image classes are represented at different levels of semantics as content and context respectively.

In this paper, we propose a Bayesian formulation to unify both local and global content-based similarities for image matching. Our proposed framework uses support vector machines to extract and combine intra-image and inter-class semantics. Support vector detectors are first trained on semantically meaningful regions and used to form detection-based image indexes. The indexes then serve as input for support vector learning of image classifiers to generate class-relative indexes. During image retrieval, similarities based on both detection-based and class-relative indexes are combined to rank images. Query-by-example experiments on 2400 home

0-7803-8554-3/04/$20.00 ©2004 IEEE.

photos with 16 semantic queries show that the combined matching approach is better than matching with single index. It also outperformed the fusion of color and texture features by 55% in average precision.

2. UNIFYING LOCAL AND GLOBAL SIMILARITIES

Given an image retrieval system, the hidden information need of a user can be modeled as the posterior probability of the set of relevant images R given the information need expressed as query q and an image x in the database, $P(R|q,x)$. The goal of the system is to return and rank images in decreasing probabilities of relevance to the user. We believe that both local (intra-image) and global (inter-class) similarities play complementary roles in image matching and ranking. From a Bayesian formulation, we have

$$P(R|q,x) = \frac{P(q,x|R) \cdot P(R)}{P(q,x)} \tag{1}$$

We observe that $P(q,x)$ tends to be small if q,x are similar (i.e. less likely to find similar images than dissimilar pair in a large database) and $P(q,x|R)$ tends to be large if q,x are similar w.r.t. R (i.e. q,x are more likely to co-occur in R if they belong to R). $P(R)$ is constant for a given query session. Hence $P(R|q,x)$ is proportional to the similarity of q,x given R (denoted as $\mu(q,x)$) and the similarity of q,x in content (denoted as $\lambda(q,x)$ i.e.

$$P(R|q,x) \propto \mu(q,x) \star \lambda(q,x). \tag{2}$$

where $\star$ is some confluence operator to combine the similarities. To realize strong semantic interpretation of image content, we propose learning and detection of semantic support regions (SSRs) in images for intra-content indexing. An innovation in our method is that no segmentation of regions is needed. Moreover, the soft detection decisions are reconciled across multiple resolutions and aggregated over spatial areas as local semantic indexes. To anchor the query context, we define semantic support classes (SSCs) as prototypical instances of the relevance class and use categorical memberships in a similarity measure.

2.1. Semantic Support Regions

In this paper, we train local support vector detectors on multi-scale block-based image regions, as inspired by multi-scale view-based object recognition framework [3], hence without a region segmentation step. Given a local image patch with feature vector z, a support vector classifier $\mathcal{S}_i$ is a detector for SSR i on z. The classification vector T for region z can be computed via the softmax function as

$$T_i(z) = \frac{\exp^{\mathcal{S}_i(z)}}{\sum_j \exp^{\mathcal{S}_j(z)}}. \tag{3}$$

As each support vector machine is regarded as an expert on a local semantic class, the outputs of $\mathcal{S}_i \ \forall i$ is set to 0 if there exist $\mathcal{S}_j, j \neq i$ that has a positive output.

As we are dealing with unconstrained home photos, we adopt color and texture features to characterize SSRs. A feature vector z has two parts, namely, a color feature vector z^c and a texture feature vector z^t. For the color feature, we compute the mean and standard deviation of each color channel (i.e. z^c has 6 dimensions). We use the YIQ color space over other color spaces as it performed better in our experiments. For the texture feature, we adopted the Gabor coefficients [1]. Similarly, the means and standard deviations of the Gabor coefficients (5 scales and 6 orientations) in an image block are computed as z^t (60 dimensions). Zero-mean normalization was applied to both the color and texture features. In this paper, we used polynomial kernels with a modified dot product similarity measure between feature vectors y and z,

$$y \cdot z = \frac{1}{2}\left(\frac{y^c \cdot z^c}{|y^c||z^c|} + \frac{y^t \cdot z^t}{|y^t||z^t|}\right) \tag{4}$$

To detect SSRs with translation and scale invariance in an image to be indexed, the image is scanned with windows of different scales, following the strategy in view-based object detection [3]. In our experiments, we progressively increase the window size from 20×20 to 60×60 at a step of 10 pixels, on a 240×360 size-normalized image. That is, after this detection step, we have 5 maps of detection.

To reconcile the detection maps across different resolutions onto a common basis, we adopt the following principle: If the most confident classification of a region at resolution r is less than that of a larger region (at resolution $r+1$) that subsumes the region, then the classification output of the region should be replaced by those of the larger region at resolution $r+1$. Using this principle, we start the reconciliation from detection map based on largest scan window (60×60) to detection map based on next-to-smallest scan window (30×30). After 4 cycles of reconciliation, the 20×20 map would have consolidated the detection decisions obtained at other resolutions.

Suppose a region Z comprises of n small equal regions with feature vectors $z_1, z_2, \cdots, z_n$. To account for the size of detected SSRs in the spatial area Z, the SSR detection vectors of the reconciled detection map is aggregated as

$$T_i(Z) = \frac{1}{n}\sum_k T_i(z_k). \tag{5}$$

For the data set and experiments reported in this paper, we designed 26 classes of SSRs (i.e. $S_i, i = 1, 2, \cdots, 26$ in Eq. (3)), organized into 8 superclasses as illustrated in Fig. 1. We cropped 554 image regions from 138 images and used 375 of them (from 105 images) as training data and

the remaining one-third for validation. Among all the kernels evaluated, those with better generalization result on the validation set are used for the indexing and retrieval tasks. A polynomial kernel with degree 2 and constant 1 produced the best result on precision and recall. Hence it was adopted in our experiments.

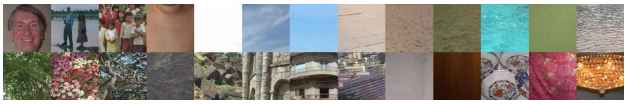

Fig. 1. Examples of semantic support regions (top-down, left-to-right): people (face, figure, crowd, skin), sky (clear, cloudy, blue), ground (floor, sand, grass), water (pool, pond, river), foliage (green, floral, branch), mountain (far, rocky), building (old, city, far), interior (wall, wooden, china, fabric, light)

For query by examples, the content-based similarity λ between a query q and an image x can be computed in terms of the similarities between their corresponding local regions. For example, the similarity based on L_1 distance measure between query q with m local regions Y_j and image x with m local regions Z_j is defined as

$$\lambda(q,x) = 1 - \frac{1}{2m}\sum_j \sum_i |T_i(Y_j) - T_i(Z_j)| \qquad (6)$$

2.2. Semantic Support Classes

As our test images are home photos, we designed a taxonomy for home photos as shown in Fig. 2. This hierarchy of categories is more comprehensive than that addressed in [5]. We trained support vector classifiers C_k on the 7 disjoint categories (SSCs) represented by the leaf nodes (except the `Miscellaneous` category) in Fig. 2 using polynomial kernels with degree 2 and constant 1. Using the softmax function, the output of classification given an image x is computed as,

$$R_k(x) = \frac{\exp^{\mathcal{C}_k(x)}}{\sum_j \exp^{\mathcal{C}_j(x)}}. \qquad (7)$$

For each class, a human subject was asked to define the list of ground truth images from the 2400 collection and 20% of the lists was used for training. To ensure unbiased training samples, we generated 10 different sets of positive training samples from the ground truth list for each class based on uniform random distribution. The negative training (test) examples for a class are the union of positive training (test) examples of the other 6 classes and the `miscellaneous` class. The classifier training for each class was carried out 10 times on these different training sets and the support vector classifier of the best run was retained. The

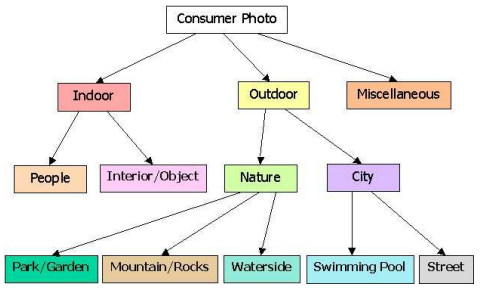

Fig. 2. Taxonomy for home photos

feature vector for classification is the image index $T_i(Z)$ as described above. To be consistent with the SSR training, we adopted the polynomial kernels with degree 2 and constant 1 with modified dot product similarity measure between image indexes $u = T_i(Y_j)$ and $v = T_i(Z_j)$ as

$$u \cdot v = \frac{1}{m}\sum_j \frac{\sum_i T_i(Y_j)T_i(Z_j)}{\sqrt{\sum_k T_k(Y_j)^2}\sqrt{\sum_k T_k(Z_j)^2}} \qquad (8)$$

The inter-class similarity μ between a query q and an image x is computed as

$$\mu(q,x) = 1 - \frac{1}{2}\sum_k |R_k(q) - R_k(x)| \qquad (9)$$

3. EMPIRICAL EVALUATION

Our test collection are 2400 genuine home photos taken over 5 years in several countries with both indoor and outdoor settings. After removing possibly noisy margins, the images are size-normalized to 240×360. Fig. 3 displays typical photos in this collection. Photos of bad quality (e.g. faded, over-exposed, blurred, dark etc) are also retained to reflect the complexity of the data. We defined 16 semantic queries and their ground truths (G.T.) among the 2400 photos (Table 1). In fact, Fig. 3 shows, in top-down left-to-right order, 2 relevant images for queries Q01-Q16 respectively. As these unconstrained home photos have highly varied and complex contents, we represent each query with 3 relevant photos as query examples in our experiments. The precisions and recalls were computed without the query images themselves in the lists of retrieved images.

Since only ranking matters for practical image retrieval and the actual relation between $\mu(q,x)$ and $\lambda(q,x)$ is unknown, we attempted linear combination ($\omega \in [0,1]$) in this paper. Replacing $P(R|q,x)$ with a combined similarity measure ρ, we have

$$\rho(q,x) = \omega \cdot \lambda(q,x) + (1-\omega) \cdot \mu(q,x) \qquad (10)$$

For query with multiple examples, $q = \{q_1, q_2, \cdots, q_K\}$, the similarity $\rho(q,x)$ is computed as

$$\rho(q,x) = max_i \rho(q_i,x) \qquad (11)$$

Fig. 3. Sample images for queries 01 to 16

Table 1. Results of QBE experiments

Query	Description	G.T.	CT	LG
Q01	indoor	994	0.62	0.91
Q02	outdoor	1218	0.78	0.91
Q03	people close-up	277	0.16	0.36
Q04	people indoor	840	0.59	0.90
Q05	interior or object	134	0.18	0.43
Q06	city scene	697	0.49	0.79
Q07	nature scene	521	0.35	0.80
Q08	at a swimming pool	52	0.18	0.57
Q09	street or roadside	645	0.50	0.81
Q10	along waterside	150	0.17	0.37
Q11	in a park or garden	304	0.71	0.81
Q12	at mountain area	67	0.28	0.24
Q13	buildings close-up	239	0.35	0.40
Q14	close up, indoor	73	0.15	0.31
Q15	small group, indoor	491	0.32	0.56
Q16	large group, indoor	45	0.29	0.26

We compare our proposed approach ($\omega = 0.5$ in Eq. (10), denoted as "LG") with the feature-based approach that combines color and texture in a linearly optimal way (denoted as "CT"). All indexing are carried out with a 4×4 grid on the images. For the color-based signature, local color histograms of b^3 ($b = 4$ to 17) number of bins in the RGB color space were computed and compared using histogram intersection. For the texture-based signature, we adopted the means and standard deviations of Gabor coeffients and the associated distance measure as reported in [1]. The Gabor coefficients were computed with 5 scales and 6 orientations. Convolution windows of 20×20 to 60×60 were attempted. The distance measures between a query and an image for the color and texture methods were normalized within $[0, 1]$ and combined linearly similar to Eq. (10). Among the relative weights attempted at 0.1 intervals, the best overall average precision of 0.38 was obtained with a dominant influence of 0.9 from the color feature (2197 bins) and 0.1 influence from the texture feature (20×20 windows).

The overall average precision for LG is 0.59, a 55% improvement over that of CT. From Table 1, our proposed approach outperformed CT in all queries except Q12 and Q16. It attained more than 50% improvement in 10 queries (Q03-10, Q14-15). From practical point of view, our method also achieved better average precision at top $20, 30, 50$ and 100 retrieved images (Table 2). Table 2 also lists the performances of individual indexing schemes by SSR and SSC to illustrate the advantage of combined matching.

Table 2. Average precisions at top images

Avg.Prec.	CT	SSR	SSC	LG
At 20	0.64	0.76	0.71	0.84
At 30	0.59	0.70	0.68	0.78
At 50	0.52	0.62	0.64	0.72
At 100	0.46	0.54	0.58	0.65
Overall	0.38	0.45	0.53	0.59

4. CONCLUSION

In this paper, we have presented an image indexing framework that combines both intra-image and inter-class semantics obtained from support vector learning. Experimental results on complex home photos confirmed that the approach is very promising when compared to linear fusion of very high dimensions of color and texture signatures. Moreover, the modular framework allows easy incorporation of better detectors and classifiers. We would validate our approach further with other image collections (e.g. medical images) in the near future.

5. REFERENCES

[1] B.S. Manjunath and W.Y. Ma. Texture features for browsing and retrieval of image data. *IEEE Trans. on PAMI*, 18(8): 837-842, 1996.

[2] M.R. Naphade and T.S. Huang. A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE Trans. on Multimedia*, 3(1): 141-151, 2001.

[3] P.C. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Proc. ICCV*, pp. 555-562, 1997.

[4] A.W.M. Smeulders et al. Content-based image retrieval at the end of the early years. *IEEE Trans. on PAMI*, 22(12): 1349-1380. 2000.

[5] A. Vailaya et al. Bayesian framework for hierarchical semantic classification of vacation images. *IEEE Trans. on Image Processing*, 10(1): 117-130, 2001.