Localized Content Based Image Retrieval

Rouhollah Rahmani, Sally A. Goldman, Hui Zhang, Sharath R. Cholleti, and Jason E. Fritts

***Abstract*— We define *localized content-based image retrieval* as a CBIR task where the user is only interested in a portion of the image, and the rest of the image is irrelevant. In this paper we present a localized CBIR system, ACCIO!, that uses labeled images in conjunction with a multiple-instance learning algorithm to first identify the desired object and weight the features accordingly, and then to rank images in the database using a similarity measure that is based upon only the relevant portions of the image. A challenge for localized CBIR is how to represent the image to capture the content. We present and compare two novel image representations, which extend traditional segmentation-based and salient point-based techniques respectively, to capture content in a localized CBIR setting.**

***Index Terms*— machine learning, content-based image retrieval, multiple instance learning, salient points**

I. INTRODUCTION

Classic *content-based image retrieval* (CBIR) takes a single *query image*, and retrieves similar images. Since the user typically does not provide any indication of which portion of the image is of interest, such a search must rely upon a global view of the image. We define *localized content-based image retrieval* as a CBIR task where the user is only interested in a portion of the image, and the rest is irrelevant.

Unless the user explicitly marks the region of interest, localized CBIR must rely on multiple images (labeled as positive or negative) to learn which portion of the image is of interest. The *query set* contains a set of images either directly provided by the user or obtained using relevance feedback [25] to add labeled feedback images to the original query image. For example, frames from surveillance video could be available for times when suspicious activity occurred (labeled as positive) and others for times when nothing out of the ordinary occurred (labeled as negative). Used in conjunction with an image repository containing unlabeled video frames, ACCIO! could be used to search for frames that have some object in common with those containing suspicious activity. In localized CBIR, the query set is used to identify the portion(s) of the image that are relevant to the user's search, as well to determine an appropriate weighting of the features.

Many CBIR systems either subdivide the image into predefined blocks [21], [22], [28], or more commonly partition the image into different meaningful regions by applying a segmentation algorithm [24], [30]. In both cases, each region of the image is represented as a vector of feature values extracted from that region. Other CBIR systems extract salient points [15], [16], [20], [27], [31], [32], which are points of high variability in the features of the local pixel neighborhood. With salient point-based methods, one feature vector is created for each salient point.

One distinction between region-based CBIR systems and localized CBIR is how the image is processed. *Single feature vector* CBIR systems represent the entire image as one feature vector. For example, a color histogram [5], [11], [18] defined over the entire image is such a representation. In contrast, *multiple feature vector* CBIR systems represent the image as a collection of feature vectors with one feature vector for either a block in some prespecified image subdivision (e.g., [21], [22]), the region defined by a segmentation algorithm (e.g., [30]), or a window around each salient point (e.g., [15], [16], [20], [27], [31], [32]).

Another important distinction is the type of similarity metric used to rank the images. In a *global ranking method*, all feature vectors in the image representation affect the ranking. While salient point-based methods only use portions of the image around the salient points, if the ranking method uses all salient points, then it is a global method. In contrast, *local ranking methods* select only portions of an image (or a subset of the salient points) as relevant to rank the images. For example, if a salient point-based method learns which subset S of the salient points are contained in desirable images and ranks images using only the subset S of salient points, then it is a local ranking method. Localized CBIR systems must use local ranking methods.

We present ACCIO!, named after Harry Potter's summoning charm, that uses a small query set in conjunction with a multiple-instance learning algorithm to identify the desired local content, re-weight the features, and then rank new images. We also present two novel image representations. The first extends traditional segmentation-based techniques, and the second extends traditional salient point-based techniques. These image representations allow ACCIO! to perform well even when the desired content is complex, defined by multiple parts or objects.

To evaluate our work, we introduce two benchmarks[1], *SIVAL*, that contains 1500 images among 25 categories, and also a benchmark composed from Flickr containing 20 categories with 100 images per category, as well as 2000 images that are not from any categories. The images for the 20 categories are obtained by searching for the following terms using Flickr's API: American flag, boat, cat, Coca-Cola can, fire flame, fireworks, honey bee, Irish flag, keyboard, Mexico City taxi, Mountie, New York taxi, orchard, ostrich, Pepsi can, Persian rug, samurai helmet, snow boarding, sushi, waterfall. The top 200 images (based on relevance) are downloaded for each category, and then we manually picked 100 images that best represented the category. For the 2000 random images we searched for the word "object," and use the top 2000 images. The specific set of images used is listed at `www.cse.wustl.edu/~sg/accio/flickr-data-set`. We use the Flickr data set to illustrate how ACCIO! can successfully be used for real image retrieval problems where the user is interested in general object categories.

This paper is organized as follows. Section II briefly discusses related work. ACCIO! and the segmentation-based representation are described in Section III. An alternative salient point-based image representation is presented in Section IV. Experimental results are presented in Section V, and we conclude and discuss future work in Section VI.

[1]Both benchmark data sets are available at `www.cse.wustl.edu/~sg/accio`.

II. Related Work

Several researchers have applied *multiple-instance* (MI) learning to localized CBIR [1], [3], [4], [21], [22], [33], [38]. Unlike standard supervised learning in which each point (instance) is labeled in the training data, in the MI model [7] each example is a *bag* of points labeled as to whether any single point within the bag is positive. The individual points are not given a label.

In this paper, we define the target concept as $(\vec{t}, \vec{s})$ for point $\vec{t} = (t_1, \ldots, t_d)$ and scale (weight) vector $\vec{s} = (s_1, \ldots, s_d)$. Similarly, each hypothesis is represented by two feature vectors, $\vec{h}$ and $\vec{s}$, where $\vec{h}$ represents the hypothesized feature vector and $\vec{s}$ is a weight vector. For an arbitrary point $\vec{p} = (p_1, \ldots, p_d)$, the weighted Euclidean distance is

$$dist_{\vec{s}}(\vec{t}, \vec{p}) = \sqrt{\sum_{i=1}^{d} (s_i(p_i - t_i))^2},$$

and the label for $\vec{p}$ is

$$\ell_{\vec{p}} = e^{-dist_{\vec{s}}(\vec{t}, \vec{p})^2}.$$

When the weighted Euclidean distance is 0, the label is 1, and as this distance approaches ∞, the label approaches 0.

In standard supervised learning, the training data is

$$D = \langle (\vec{p}_1, \ell_{p_1}), (\vec{p}_2, \ell_{\vec{p}_2}), \ldots, (\vec{p}_m, \ell_{\vec{p}_m}) \rangle,$$

where $\ell_{\vec{p}_i}$ is the label for $\vec{p}_i$. For example, consider when a color histogram is used to represent an image as a single point. Target $\vec{t}$ is the feature vector representing the ideal color histogram, and $\vec{s}$ is the ideal weighting of the features. This approach captures how machine learning approaches are often applied in CBIR to use training data from relevance feedback to re-weight features.

In MI learning the examples are bags of points. More formally, the training data is $D = \{\langle B_1, \ell_1 \rangle, \langle B_2, \ell_2 \rangle, \ldots \langle B_m, \ell_m \rangle\}$ where $B_i = \{\vec{p}_{i,1}, \vec{p}_{i,2}, \ldots \vec{p}_{i,|B_i|}\}$. Let $\ell_{i,j}$ be the label of point $\vec{p}_{i,j} \in B_i$. Then $\ell_i = \max\{\ell_{i,1}, \ell_{i,2}, \ldots, \ell_{i,|B_i|}\}$. When each bag contains a single point, MI learning reduces to standard supervised learning. CBIR systems that use MI learning associate a bag with each image, and each point in the bag is a feature vector representing either (1) a fixed region of the image, (2) a segment of the image, or (3) the window around a salient point.

Chen and Wang [4] and Bi et al. [3] considered the problem of image categorization. For this problem, one would typically have fairly large training sets for each image category. A related, though different task, is to use a large set of images, each annotated with several key words, to learn a correspondence between objects and words (e.g., [2], [8]).

Salient point-based representations are also commonly used in CBIR [15], [16], [20], [27], [31], [32]. In general, such methods rank images based on the number of salient points that match between a single query image and the images in the repository. Such methods are global in that features from all the salient points in the image are used in the ranking.

Scale invariant feature transform (SIFT) [20] was introduced as a way to extract salient points that are invariant to many common image transforms. Mikolajczyk and Schmid [23], [26] compared a variety of approaches of identifying salient points, and found SIFT to work best for image matching. SIFT and variations of it have also been used for image retrieval [15], [16], [31], [32]. Some prior work has combined image segmentation and salient points [17]. However, it uses global ranking methods, which are not well-suited for localized CBIR.

III. Accio! - A New Localized CBIR System

In the Accio! CBIR system, the user provides a query set either directly or by adding the feedback set to the original query image. The Accio! system first segments the image and converts the segmented image into the MI representation using the *bag generator*. Then Accio! uses *generalized EM-DD*[2] (GEM-DD), a generalized version of the EM-DD algorithm [37]. GEM-DD performs a gradient search with multiple starting points to obtain an ensemble of hypotheses that are consistent with both the positive and negative images in the query set. Finally, a ranking algorithm combines the hypotheses from the learning algorithm to obtain an overall ranking of all images in the repository. The user can then label some of the ranked images as "desired" or "not desired' to augment the query set.

A. Our Segmentation-Based Image Representation

We now present our new image representation, *segmentation with neighbors*, which combines the robustness of segmentation with the contextual awareness of neighbors. We first transform all images into the YCrCb color space[3] and use a wavelet texture filter so that each pixel in the image has three color features and three texture features [36]. Alternate features could be used as desired. Next, the IHS segmentation algorithm [35] is used to segment the image. A different segmentation algorithm could likewise be used instead.

Since often it is the immediate surroundings that allow for a more complex and descriptive definition of the content of the image, we compute the neighbors to the north, south, east, and west for every segment. The feature vector for each segment is augmented with the feature differences between its features and its neighbors' features for all four neighbors. We use the feature differences to allow for robustness against global changes in the image, such as brightness changes from variable light or shade.

We view each segment x in image I as a 30-dimensional point where the first six features are the average color and texture values for x. The next six features hold the difference between the average color and texture values of the northern neighbor and x. Similarly there are six features for the difference information between x and each of its other three cardinal neighbors. The query set has one bag for each image in the query set.

B. The GEM-DD Algorithm

EM-DD [37] treats the knowledge of which point corresponds to bag's label as a missing attribute and applies the *expectation-maximization* (EM) algorithm [6] to convert the MI learning problem to a standard supervised learning problem. It starts with an initial value for target point $\vec{h}$ and scale vector $\vec{s}$, and then repeatedly performs the following two steps. In the first step (E-step), the current $\vec{h}$ and $\vec{s}$ are used to pick one point from each bag that is most likely (given the generative model) to be the one responsible for the label. In the second step (M-step), a two-phase gradient search is used to find the $\vec{h}$ and $\vec{s}$ that maximizes the DD measure[4].

[2]The conference version [24] used Ensemble EM-DD, which has since been improved to obtain the GEM-DD algorithm.

[3]Accio! could be easily modified to use a different color space, if desired.

[4]Technically, the negative log of the diverse density is minimized.

We now describe our new MI algorithm, GEM-DD, which is built upon EM-DD. It starts at an initial point from a randomly selected set of positive bags, with different initial scale factors used to weight the given segment relative to its neighbors. Specifically, the initial weights used for the segment itself (versus its neighbors) are 100%, 80%, 60%, 20% (all 5 regions, equally weighted), and 0%, with the remaining percent equally divided among the neighboring regions. Second, all initial scale factors are adjusted based on the characteristics of the training data and the floating point precision of the computing platform. Third, GEM-DD performs feature normalization so all features are treated equally when weighted equally. In particular it uses the *training data*, normalizing the range of features values into the range 0 to 1. The same normalization factors are applied to the test data.

Finally, GEM-DD returns a set of hypotheses that help provide several independent ways to characterize the desirable images. Let $\mathcal{H} = \{(\vec{h}_1, \vec{s}_1), \ldots, (\vec{h}_k, \vec{s}_k)\}$ be the set of hypotheses and associated scales returned by GEM-DD, sorted in descending order by their DD measure. Let I be an image in the image repository that is segmented into r segments and represented by the bag $B_I = \{\vec{p}_1, \ldots, \vec{p}_r\}$. The Hausdorff distance between hypothesis $H_i = (\vec{h}_i, \vec{s}_i)$ and bag B_I is given by

$$d(H_i, B_I) = \min_{j=1,\ldots,r} dist_{\vec{s}_i}(\vec{h}_i, \vec{p}_j)$$

where $dist$ is the weighted Euclidean distance.

DD and EM-DD use the **minNLDD** measure to rank bags according to the "best" hypothesis H_1. Zhang et al. [38] introduced **AvgAll**, which ranks test bag B according to the average label

$$\ell_B = \frac{1}{k} \cdot \sum_{i=1}^{k} e^{-d(H_i,B)^2}.$$

While one would expect that there are a set of hypotheses that provide independent ways to characterize the images of interest to the user, there are also some hypotheses that result from a bad starting point for EM. Furthermore, one would expect that hypotheses with a high DD value are "good" hypotheses while the hypotheses with low DD values should be excluded. Thus, we parameterize GEM-DD by an integer τ where $1 \leq \tau \leq k$, and label bag B according to

$$\ell_B = \frac{1}{\tau} \cdot \sum_{i=1}^{\tau} e^{-d(H_i,B)^2}.$$

Setting $\tau = 1$ gives minNLDD, and setting $\tau = k$ gives AvgAll. Finally, the images are ranked in decreasing order based on the ℓ_B values.

IV. Our Novel Salient Point-Based Image Representation:

Salient point-based representations decouple the sensitivity of a CBIR system from the quality of the segmentation. Traditional uses of salient points for CBIR compute the feature vector for a salient point according to the features of all pixels in a window around the salient point [15], [16], [20], [27], [31], [32]. However, since salient points are often on the boundary of objects, the features assigned to a salient point often involve pixels from different objects. While this is acceptable in standard (global) CBIR systems, which use all portions of the image for retrieval, for localized CBIR it is crucial to find a good representation for individual segments that faithfully represents local regions. Another drawback of using traditional salient point-based extraction methods is these points often gather at more textured areas, so many salient points capture the same portions of the image.

We introduce a new salient point representation for localized CBIR that is achieved using two orthogonal techniques. First, we use image segmentation to form a mask that limits the number of salient points in each segment while maintaining the diversity of the salient points. Second, we use the local characteristics of the pixel window around a salient point to determine how to split the window into two sub-windows, and assign each sub-window features based on both it and its neighboring sub-window. We now describe these two methods in more depth.

A. *SPARSE* (Salient Points Auto-Reduction using SEgmentation)

Our SPARSE image representation limits the number of salient points in each segment while maintaining the diversity needed for localized CBIR. SPARSE first applies a salient point detection algorithm to the image. We use a Harr wavelet-based salient point detection method. Beginning at the coarsest level of wavelet coefficients, we keep track of the salient points from level to level by finding the points with the highest coefficients on the next finer level among those used to compute the wavelet coefficients at the current level. The saliency value of a salient point is the sum of the wavelet coefficients of its parent salient points from all coarser scales. Note, any salient point detection method can be used here instead, with little modification.

Next a segmentation algorithm is applied to the image. The segmentation algorithm we use is a clustering-based segmentation method [10] that uses the Euclidean distance between 6-dimensional feature vectors, with 3 color features and 3 texture features, as its similarity measure. The resulting segmentation is used to reduce the number of salient points. Specifically, SPARSE keeps at most k salient points in each segment, by keeping those with the highest saliency value. In our implementation, $k = 3$.

Fig. 1 shows examples of salient points detected using SPARSE. For comparison, we also show the salient points detected by the Harr wavelet-based salient points detection method, and the SIFT [20] method. The wavelet-based method selects the top 200 salient points for each image. SPARSE reduces it to at most 96 salient points per image. SIFT selects 392 salient points for the tea box, and 288 salient points for the coke can. When using SPARSE the salient points predominantly gather at complex objects, whereas with the wavelet-based method the salient points gather at the edges. While the wavelet-based method does reduce the number of salient points on the textured region (such as at the printed words on the calendar and tea box), SPARSE further reduces the number of salient points at textured regions.

B. *VSWN: Our Salient Point Representation*

We now describe our *variably-split window with neighbor* (VSWN) representation. Since salient points are often on the boundary of objects, the features assigned to a salient point involve pixels from different objects, which is not good for localized CBIR because only one of these objects might be of interest. If we divide the window, we can better capture the color and texture of an individual object.

For each salient point, VSWN uses the local characteristics of the window around each salient point to split the window in either

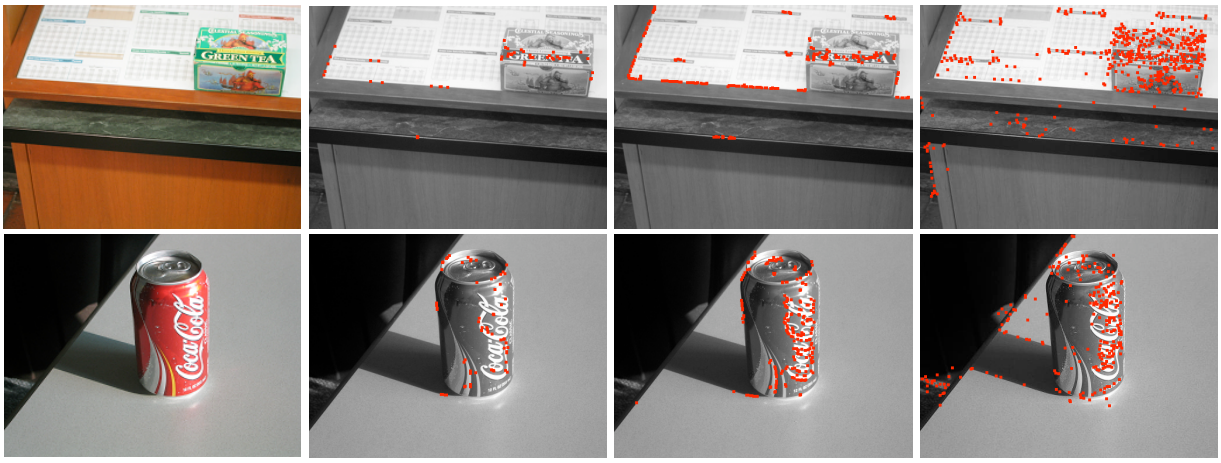

Fig. 1. Salient points detection with SPARSE, the Harr Wavelet-based method, and SIFT.

the horizontal, vertical, or one of the two diagonal directions. The VSW (*variably split window*) technique adaptively chooses the best split. VSW applies a wavelet transform on the pixels in the window, and measures the average coefficients in the HL (vertical), LH (horizontal), and HH (up-right diagonal). We also flip the window around the vertical to compute a flipped-HH coefficient (up-left diagonal). If the LH and HL channels have similar coefficients, then we use the split associated with the larger of the HH and flipped-HH channel. Otherwise, we use the split based on the largest of the four channels. While the best segmentation of the region is unlikely to be one of the four splits considered (since it is an 8x8 window) the selected split serves as a sufficiently good approximation. If desired, we could further subdivide each sub-window.

Second, as was the case for segmentation with neighbors, for salient points it is advantageous to incorporate information about the neighboring sub-window to provide additional context. The two sub-windows for each salient point are represented via three color features and three texture features. VSWN augments the feature vector for each sub-window with the *difference* between its values and the other sub-window's values for each of the six features. We use the feature differences to allow for robustness against global changes in the image, such as brightness changes from variable light or shade. Since we do not know which sub-window might hold the object of interest, we create two 12-dimensional feature vectors for each salient point: one for each sub-window as the object of interest.

V. Experimental Results

We compare the system-level performance of Accio!, SIMPLIcity [14], [30], and SBN [22] on the SIVAL and a COREL natural scenes data sets, both with small query sets (2-16 images), for which Accio! was designed, and with the traditional CBIR setting of a single positive query image. We also compare the performance of our SPARSE+VSWN salient point-based representation to that of SIFT and the wavelet-based method. On the Flickr data set we compare our segmentation-based and salient point-based representations. Unless otherwise indicated, Accio! results were produced using the segmentation-based representation, where τ was set roughly equal to the bag size. Thus, $\tau = 25$ for the segmentation-based representation, and $\tau = 75$ for the salient point-based representations. For results that used a single query image, we set $\tau = |\mathcal{H}|$.

For the SBN algorithm we replace DD by the EM-DD algorithm because of its performance gains in both retrieval accuracy and efficiency [38]. Since SIMPLIcity is designed to use a single positive example, we created a variant of it that uses any size query image set. Let $\mathcal{P}$ be the set of positive images, and let $\mathcal{N}$ be the set of negative images. For image q in the query set and image x in the image repository, let $r_q(x)$ be the ranking SIMPLIcity gives to image x when the query image is q. (The highest rank image is rank 0.) Our variation of SIMPLIcity ranks the images in decreasing order based on

$$\prod_{q\in\mathcal{P}}\left(1-\frac{r_q(x)}{n}\right)\cdot\prod_{q\in\mathcal{N}}\frac{r_q(x)}{n}.$$

We selected this measure since it is similar to the definition of diverse density of a point t, $DD(t) = \prod_{q\in\mathcal{P}\cup\mathcal{N}} \Pr(t|q)$. For an image $q \in \mathcal{P}$, $(1 - r_q(x)/n)$ can be viewed as the probability that x is positive given that q is positive. Similarly, for an image $q \in \mathcal{N}$, $r_q(x)/n$ can be viewed as the probability that x is positive given that q is negative. When given a single positive image, the ranking is the same as that given by the original SIMPLIcity algorithm [14], [30].

As our measure of performance, we use the area under the ROC curve [12] that plots the true positive rate as a function of the false positive rate. The area under the ROC curve (AUC) is equivalent to the probability that a randomly chosen positive image will be ranked higher than a randomly chosen negative image. Unlike the precision-recall curve, the ROC curve is insensitive to the ratio of positive to negative examples in the image repository. Regardless of the fraction of the images that are positive, for a random permutation the AUC is 0.5. For all AUCs reported, we repeat each experiment with 30 random selections of the positive and negative examples and use these to compute the average AUC and the 95% confidence intervals for the AUC.

A. System Performance

Table I compares the average performance (over all categories) of Accio!, SIMPLIcity and SBN for the SIVAL and the natural scenes data sets. Fig. 2 compares all 25 object categories of

	SIVAL		Natural Scenes	
system	8 pos, 8 neg	Single pos	8 pos, 8 neg	Single Pos
Accio!	81.8	53.5	83.6	67.2
Accio! (SPARSE+VSWN)	81.6	56.3	-	-
SIMPLIcity	57.9	55.7	74.8	73.7
SBN	53.9	50.3	73.6	61.4
Accio! (conference version)	74.6	61.0	87.7	74.5

TABLE I
SUMMARY OF AUC VALUES AVERAGED OVER THE CATEGORIES OF SIVAL AND COREL NATURAL SCENES DATASETS.

SIVAL when the query set contains 8 random positive and 8 random negative examples. For 2 categories - "LargeSpoon" and "CandleWithHolder" - SIMPLIcity's segmentation algorithm failed on a few images, so results could not be provided. For both representations, in every category ACCIO! 's performance is statistically better than that of both SIMPLIcity and SBN, with the exception of "LargeSpoon" for SBN, and "RapBook" for SIMPLIcity. ACCIO! using segmentation with neighbors has an average improvement of 51.7% over SIMPLIcity and 41.4% over SBN. ACCIO! using SPARSE+VSWN has an average improvement of 51.2% over SIMPLIcity and 41.0% over SBN.

Fig. 2 also compares the SPARSE+VSWN and the segmentation with neighbors representation of ACCIO! . We see similar performance in 17 of 25 categories. In 5 categories, the segmentation-based approach performs statistically better, and in 3 categories, SPARSE+VSWN is statistically better. So overall, their results are comparable. Segmentation with neighbors encodes its neighbors in a manner that preserves orientation which is advantageous for some tasks (e.g., when distinguishing a waterfall from a river). Furthermore, encoding four neighbors, instead of just one, captures more contextual information for each segment. On the other hand, SPARSE+VSWN has several advantages over segmentation with neighbors. The reduction in dimensionality from 30 to 12 improves the time complexity. Also VSWN's use of a single neighbor that is both mirror invariant and rotation invariant, allows it to perform better on categories in which the images experience significant rotation (90° and 180°). While the salient points currently encode only the same information as the segmentation-based method, salient points can be encoded with a multitude of additional features not easily derived from segmentation methods, such as the orientation histogram used by SIFT. Additionally, salient points by their nature can capture much finer detail in the image than segmentation.

Since SIMPLIcity was designed for a single positive query image, we also considered when the query set contains only a single positive image (not shown). On average ACCIO! obtains a 4.3% improvement in performance over SIMPLIcity, and a 15.6% improvement over SBN. For our alternative SPARSE+VSWN representation, ACCIO! obtains a 12.4% improvement in performance over SIMPLIcity, and a 24.5% improvement over SBN. The version of SIMPLIcity we created to make use of a query set with multiple images did improve performance over having a single query image in 12 of the 23 categories for which we obtained data.

Fig. 3 shows the performance of ACCIO! when using the segmentation with neighbors representation on the Flickr data set for varying query set sizes. Likewise, Fig. 4 shows the performance of ACCIO! when using SPARSE+VSWN. As the size of the training data increases, in general we both get better retrieval performance, and also the variation in performance is reduced. For some categories increasing the training size has a very small impact (e.g., waterfall, samurai helmet), yet for others (e.g., American flag, Pepsi can) the impact is quite large. When there is some aspect of a category that is very distinctive then the smaller training set can be effective. However, when the object is defined by fairly typically occurring colors/textures (e.g., the color red) or in a category with a lot of variation (e.g., fire flame), having a larger query set can really help performance.

We now compare the performance of these two representations when there are 8 positive and 8 negative examples. ACCIO! using segmentation with neighbors performs statistically better on "Snowboarding," "Sushi," and "Persian Rug." Conversely, though ACCIO! using SPARSE+VSWN does not perform statistically better in any of the image categories, there is a noticeable improvement over segmentation with neighbors on the "American Flag," "Fire Flame," "Pepsi Can," and "Coca-Cola Can" categories. Since the "American Flag," "Pepsi Can," and "Coca-Cola Can" images all contain a specific complex object of interest, there will be a large number of salient points in each image corresponding to that object. Since the object is specific, the colors and textures defining the object are very well defined, with variations due only to lighting, shading, and noise. Therefore, the sets of salient points for these objects serve as effective identifiers enabling the SPARSE+VSWN method to perform well on such categories. The "Fire Flame" category, does not target as specific an object, but the nature of fire does lend itself well to a small set of easily distinguishable colors and textures.

The "Snowboarding," "Sushi," and "Persian Rug" categories do not contain a specific object, but include objects with common sets of color schemes and textures, which can be effectively captures by whole-segment feature characteristics. As long as the segmentation algorithm can effectively segment out the object regions, and there are sufficient training examples to characterize both the most common variations in the object's color and texture and the spatial relationship between the object(s) of interest, ACCIO! using segmentation with neighbors generally performs well. For example, in "Snowboarding", segmentation with neighbors searches for regions containing snow bordered by one or more neighboring regions containing trees, mountains, or sky. While these categories can be characterized by a variety of different color schemes and textures, with effective segmentation and sufficient training data, segmentation with neighbors achieves good performance for these categories.

B. Comparison of Salient Point-Based Representations

In this section we compare the performance of ACCIO! when using salient point-based representations. Fig. 5 compares the three salient point extraction methods: Wavelet, SIFT,

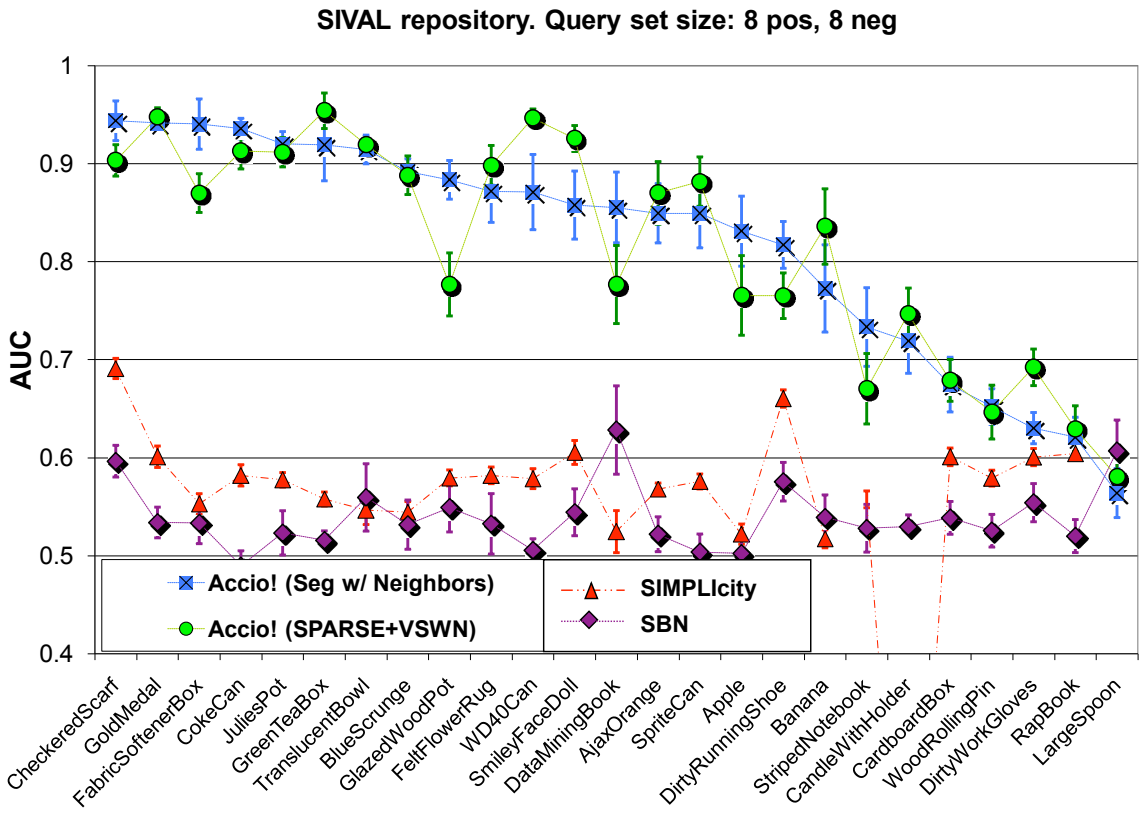

Fig. 2. CBIR systems results for the SIVAL data set when the query set has 8 positive and 8 negative examples.

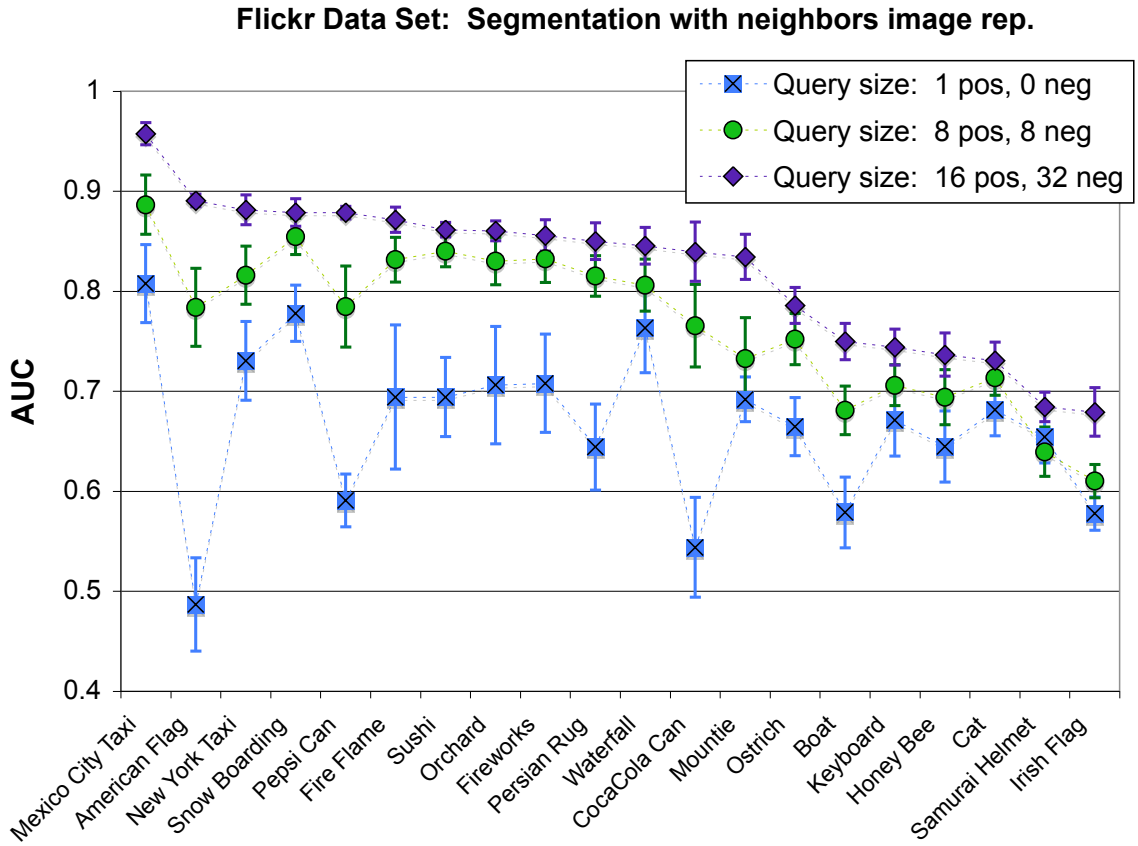

Fig. 3. Results from our segmentation with neighbors representation on the Flickr data set.

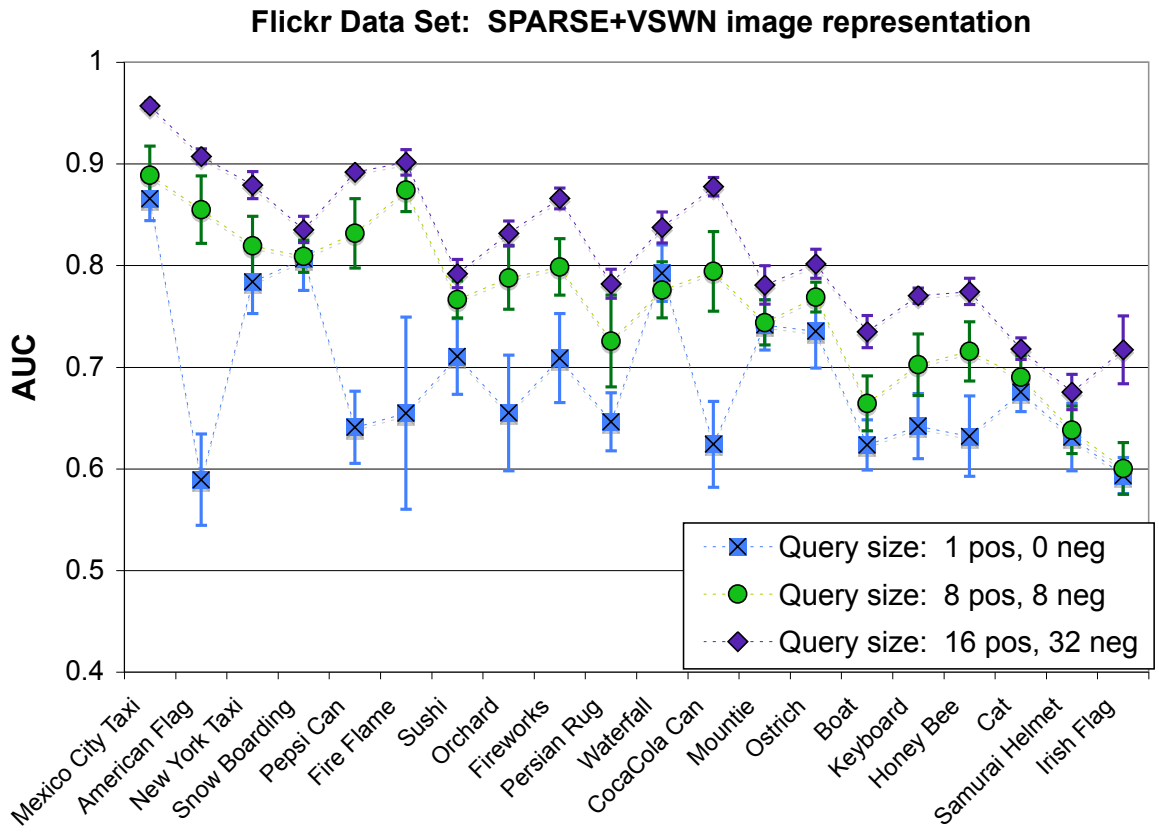

Fig. 4. Results from our SPARSE+VSWN representation on the Flickr data set.

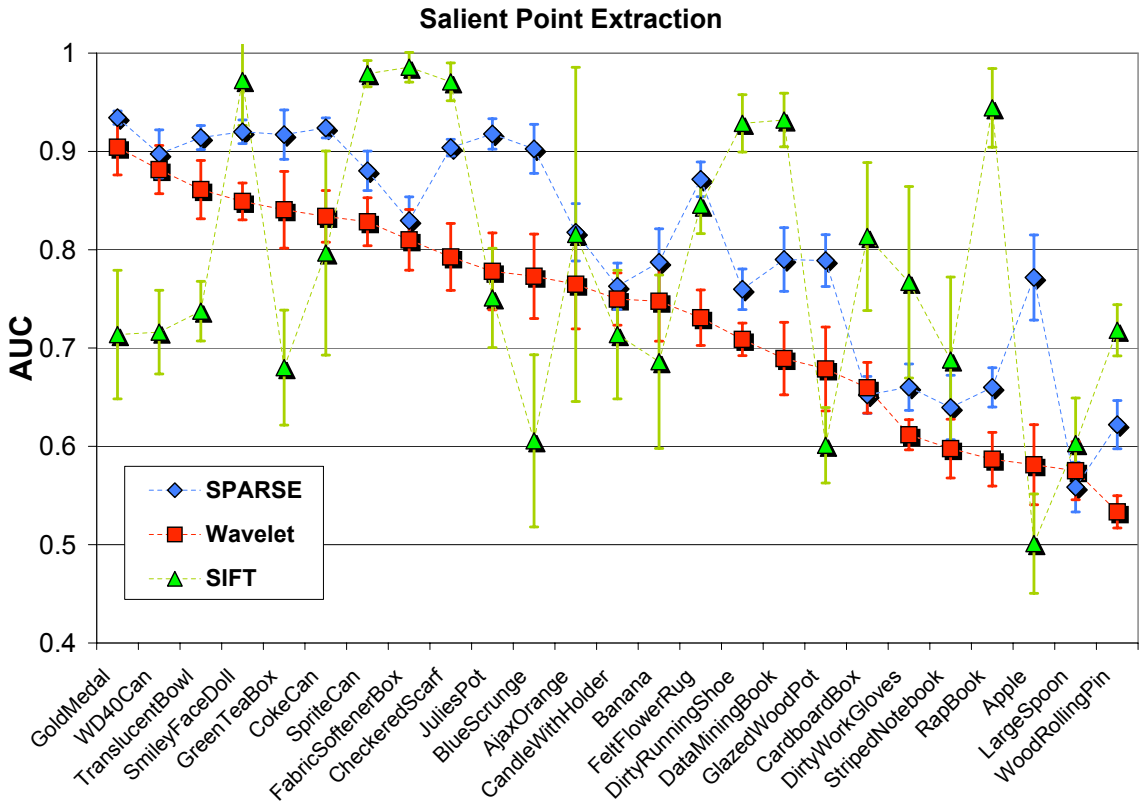

Fig. 5. Comparing salient point methods on the SIVAL data set where the query set has 8 positive and 8 negative examples.

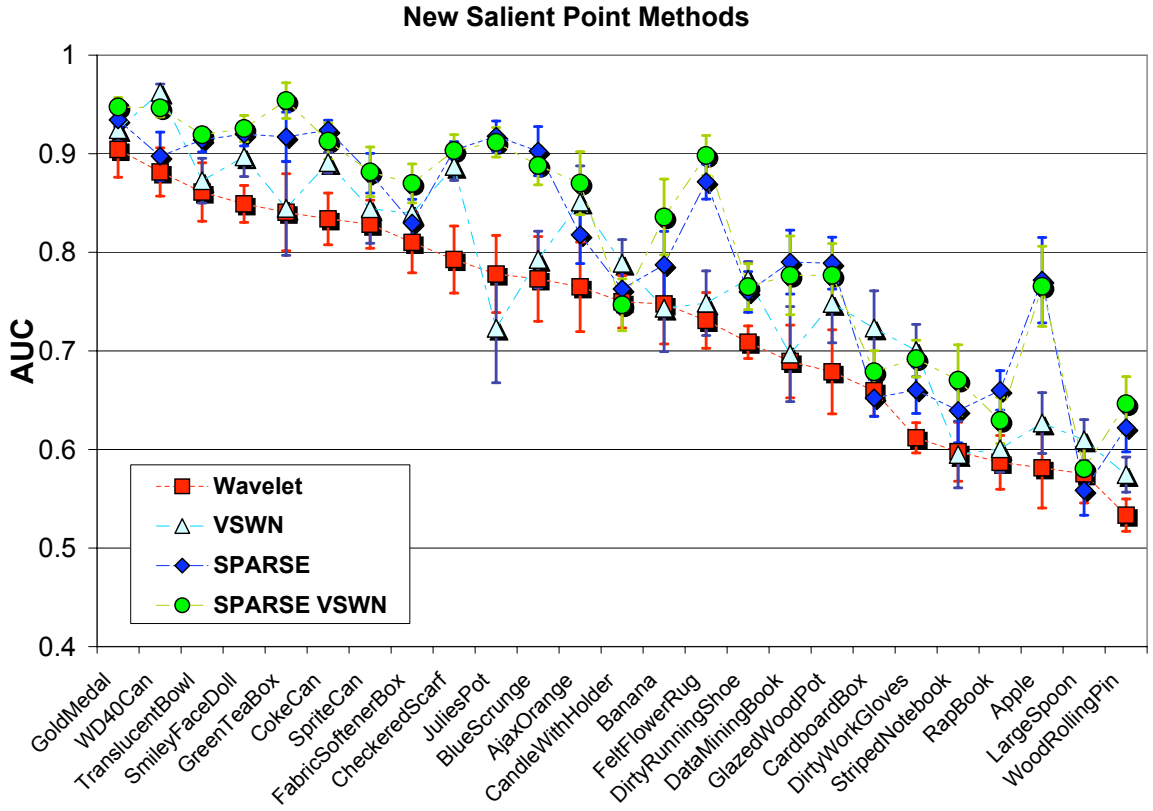

Fig. 6. Comparing salient point methods on SIVAL data set for a query set of 8 positive and 8 negative examples.

and SPARSE+VSWN. In these experiments, both SPARSE and Wavelet use the same salient point extraction and representation methods (3 color and 3 texture dimensions). The primary difference between them is where the salient points are placed in the image. SIFT both uses a different feature extraction method, placing the salient points differently, and uses a more complex feature representation. SPARSE outperforms Wavelet in 23 of 25 categories, 16 of which are statistically significant. The use of SPARSE can also improve the algorithm efficiency by reducing the number of feature vectors per bag, and hence the number of computations.

The SIFT feature vector has 128 dimensions that describe the local gradient orientation histogram around a salient point. Results using SIFT were generated using 5 random selections of the training data (as opposed to 30) since the high dimensionality makes it very computationally intensive. SIFT performs 5.9% better than Wavelet over all categories. However, SPARSE outperforms SIFT by 3.2% (over all categories), despite its relatively simpler feature representation.

We also compared the performance obtained from varying the salient point extraction method. The average AUC values (across all categories) of SIVAL are 81.6 for SPARSE+VSWN, 80.3 for SPARSE, 77.0 for VSWN, 73.5 for Wavelet, and 77.9 for SIFT. Both SPARSE and VSWN can help improve retrieval performance, and when used together they improve performance further.

Fig. 6 independently compares the effect of using SPARSE and VSWN to the standard wavelet-based salient points on SIVAL. Adding VSWN to Wavelet leads to a 4.8% improvement when averaged over all categories with statistically significant improvements in nine categories. Adding SPARSE to Wavelet leads to a 9.3% improvement when averaged over all categories. When both SPARSE and VSWN are added to Wavelet, an 11.0% increase in performance occurs.

VI. Conclusions and Future Work

We have presented Accio!, a localized CBIR system that does not assume that the desired object(s) are in a fixed location or have a fixed size. We have demonstrated that Accio! outperforms existing systems for localized CBIR on both a natural scenes image repository and SIVAL our new benchmark data set. Our experimental results when using the Flickr data set, demonstrate that Accio! can successfully be used for real image retrieval problems where the user is interested in general object categories.

We introduce the SPARSE technique, which uses segmentation as a filter to reduce the total number of salient points while still maintaining diversity. Finally, we introduce the VSWN salient point representation, which splits salient points on region boundaries into two salient points, characterizing the separate objects at the boundary.

There are many directions for future work. We believe that Accio! can be improved further by making further improvements to GEM-DD, by employing improved segmentation algorithms, and perhaps by the careful introduction of some features to represent shape. For SPARSE, an important area of future work is to perform experiments to determine the sensitivity to k, the number of salient points per segment, and develop methods to select the best value for k.

ACKNOWLEDGMENT

We would like to thank John Krettek, Ibrahim Noorzaie, and Ian Bushner for all of their valuable feedback and ideas. This material is based upon work supported by the National Science Foundation under Grant No. 0329241.

REFERENCES

[1] S. Andrews, T. Hofmann, and I. Tsochantaridis. Multiple instance learning with generalized support vector machines. *Artificial Intelligence*, pages 943–944, 2002.

[2] K. Barnard, P. Duygulu, N. de Freitas, and D. Forsyth. Matching words and pictures. *Journal of Machine Learning Research*, volume 3, pages 1107–1135, 2003.

[3] J. Bi, Y. Chen, and J. Wang. A sparse support vector machine approach to region-based image categorization. *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1121–1128, 2005.

[4] Y. Chen and J. Wang. Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research*, pages 913–939, 2004.

[5] I. Cox, M. Miller, S. Omohundro, and P. Yianilos. PicHunter: Bayesian relevance feedback. *International Conference on Pattern Recognition*, pages 361–369, 1996.

[6] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistics Society*, volume 39, pages 1–38, 1977.

[7] T. Dietterich, R. Lathrop, and T. Lozano-Pérez. Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence*, volume 89(1–2), pages 31–37, 1997.

[8] P. Duygulu, K. Barnard, N. de Freitas, and D. Forsyth. Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. *Proceedings of the 7th European Conference on Computer Vision*, pages IV:97–12, 2002.

[9] M. Fischler and R. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, volume 24, pages 381–395, 1981.

[10] D. Forsyth and J. Ponce. Computer Vision: A Modern Approach. Prentice Hall, 2003.

[11] T. Gevers and A. Smeulders. Image search engines: An overview. In G. Medioni and S. B. K. (Eds.), editors, *Emerging Topics in Computer Vision*. Prentice Hall, 2004.

[12] J. Hanley and B. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve, *Radiology*, volume 143(1), pages 29–36, 1982.

[13] X. Huang, S.-C. Chen, M.-L. Shyu, and C. Zhang. User concept pattern discovery using relevance feedback and multiple instance learning for content-based image retrieval. *Proceedings of the 8th International Conference on Knowledge Discovery and Data Mining*, pages 100–108, 2002.

[14] J. Wang J. Li and G. Wiederhold. IRM: Integrated region matching for image retrieval. *Proceedings of the 8th ACM International Conference on Multimedia*, pages 147–156, 2000.

[15] Y. Ke and R. Sukthankar. PCA-SIFT: A more distinctive representation for local image descriptors. *Proceedings of the 2004 IEEE conference on Computer Vision and Pattern Recognition*, volume 2, pages 506–513, 2004.

[16] L. Ledwich and S. Williams. Reduced sift features for image retrieval and indoor localisation. *Proceedings of Australasian Conference on Robotics and Automation*, 2004.

[17] H.-K. Lee and Y.-S. Ho. A region-based image retrieval system using salient point extraction and image segmentation. *Lecture Notes in Computer Science: Advances in Multimedia Information Processing - PCM 2002: Third IEEE Pacific Rim Conference on Multimedia*, pages 209–216, 2002.

[18] F. Long, H. Zhang, and D. Feng. Fundamentals of content-based image retrieval. In D. Feng, W. Siu, and H. Z. (Eds.), editors, *Multimedia Information Retrieval and Management- Technological Fundamentals and Applications*. Springer, 2003.

[19] E. Loupias. Salient points detection using wavelet transform, http://telesun.insa-lyon.fr/ loupias/points/demo.html.

[20] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, volume 60(2), pages 91–110, 2004.

[21] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. *Proceedings of the 1997 conference on Advances in Neural Information Processing Systems 10*, pages 570–576, 1998.

[22] O. Maron and A. Ratan. Multiple-instance learning for natural scene classification. *Proceedings of the 15th International Conference on Machine Learning*, pages 341–349, 1998.

[23] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 257–264, 2003.

[24] R. Rahmani, S. Goldman, H. Zhang, J. Krettek, and J. Fritts. Localized content-based image retrieval. *Proceedings of the 7th ACM SIGMM International Workshop on Multimedia Information Retrieval*, pages 227–236, 2005.

[25] Y. Rui, T. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool in interactive content-based image retrieval. *IEEE Transactions Circuits and Systems for Video Technology*, volume 8(5), pages 644–655, 1998.

[26] C. Schmid, R. Mohr, and C. Bauckhage. Evaluation of interest point detectors. *International Journal of Computer Vision*, volume 37(2), pages 151–172, 2000.

[27] Q. Tian, N. Sebe, M. S. Lew, E. Loupias, and T. S. Huang. Image retrieval using wavelet-based salient points. *Journal of Electronic Imaging, Special Issue on Storage and Retrieval of Digital Media*, volume 10(4), pages 935–849, 2001.

[28] Q. Tian, Y. Wu, and T. Huang. Combine user defined region-of-interest and spatial layout for image retrieval. *Proceedings of the IEEE Conference on Image Processing*, volume 3, pages 746–749, 2000.

[29] P. Torr and A. Zisserman. Mlesac: a new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding*, volume 78(1), pages 138–156, 2000.

[30] J. Wang, J. Li, and G. Wiederhold. SIMPLIcity: semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 947–963, 2001.

[31] J. Wang, H. Zha, and R. Cipolla. Combining interest points and edges for content-based image retrieval. *Proceedings of the IEEE International Conference on Image Processing*, pages 1256–1259, 2005.

[32] C. Wolf, J.-M. Jolion, W. Kropatsch, and H. Bischof. Content based image retrieval using interest points and texture features. *Proceedings of the 15th IEEE International Conference on Pattern Recognition*, volume 4, page 4234, 2000.

[33] C. Yang and T. Lozano-Pérez. Image database retrieval with multiple instance techniques. *Proceedings of the 16th International Conference on Data Engineering*, pages 233–243, 2000.

[34] H. Zhang, R. Rahmani, S. Cholleti, and S. Goldman. Local Image Representations using Pruned Salient Points with Applications to CBIR. *Proceedings of the 14th ACM International Conference on Multimedia*, pages 287-296, 2006.

[35] H. Zhang, J. Fritts, and S. Goldman. An improved fine-grain hierarchical method of image segmentation. Technical report, Washington University in St Louis, 2005.

[36] H. Zhang, J. Fritts, and S. Goldman. A fast texture feature extraction method for region-based image segmentation. *Proceedings of IS&T/SPIE's 16th Annual Symposium on Image and Video Communication and Processing*, SPIE Vol. 5685, pages 957–968, 2005.

[37] Q. Zhang and S. Goldman. EM-DD: an improved multiple-instance learning technique. *Neural Information Processing Systems*, pages 1073–1080, 2001.

[38] Q. Zhang, S. Goldman, W. Yu, and J. Fritts. Content-based image retrieval using multiple instance learning. *Proceedings of the 19th International Conference on Machine Learning*, pages 682–689, 2002.

[39] Z. Zhou and M. Zhang. Ensembles of multi-instance learners. *Proceedings of the 14th European Conference on Machine Learning*, pages 492–502, 2003.

Rouhollah Rahmani is currently serving as a post-doctoral research fellow in the Department of Bio-Electrical Engineering in the University of Tehran. He earned both a Bachelor of Arts in Mathematics and Bachelor of Science in Computer Science from Washington University in St Louis. He is expected to completed is PhD from the Department of Computer Science at Washington University in St Louis in May 2008. His primary research area is in multiple-instance learning, a subdiscipline of machine learning. In addition to developing new algorithms for multiple-instance learning, this research focused heavily on the application of image search and object extraction for both the general and medical domains. Additional research projects have included developing algorithms for finding protein-protein binding pathways,developing continuous state spaces for reinforcement learning, and working on MPEG-2 compression for Internet-2.0. He has published articles in ICML, ACM Multimedia, ACM MIR, IEEE ICTAI. He has served as a reviewer for IEEE Transactions on Neural Networks, the Journal of Machine Learning Research, and the Journal of Neurocomputing.

Sally A. Goldman is currently the Edwin H. Murty Professor of Engineering at Washington University in St. Louis where she is the Associate Chair of the Department of Computer Science and Engineering. She received a Sc.B. from Brown University and a M.S. and Ph.D. from Massachusetts Institute of Technology under the supervision of Ronald Rivest. Her primary research interests are machine learning, content-based image retrieval, and computational learning theory. She was a recipient of the NSF National Young Investigator award. Dr. Goldman's work has appeared in many top conferences including FOCS, STOC, ICML, CVPR, ACM Multimedia, NIPS, COLT, and in journals such as Journal of the ACM, SIAM Journal on Computing, Information and Computation, Journal of Computer and System Sciences, Journal of Machine Learning Research, and the Machine Learning Journal. She is currently on the editorial board for the Journal of Machine Learning Research and the Journal of Computer and Systems Sciences. She has recently published *A Practical Guide to Data Structures and Algorithms Using Java* with her husband and co-author Ken Goldman.

Hui Zhang received his B.S. in Electrical Engineering from Wuhan University, China in 1999, and a M.S. and Ph.D. in Computer Science from Washington University in St. Louis in 2003 and 2007, respectively. His areas of interest include computer vision, image processing and video processing. In particular, he has worked on low-level and middle-level image analysis, image segmentation, segmentation evaluation and salient points of images. His work has appeared in journals and conferences including CVIU, CVPR and ACM Multimedia.

Sharath R. Cholleti is a Ph.D. student at Washington University in St. Louis under the supervision of Dr. Sally Goldman in the Department of Computer Science and Engineering. He received a B.Tech. from Indian Institute of Technology, Guwahati and a M.S. from Washington University in St. Louis. His primary research interests are machine learning and content-based image retrieval. His work has appeared in ACM Multimedia, ICTAI, CVPR, and LCTES.

Jason E. Fritts received the B.S. degree from Washington University, St. Louis, MO, in 1994, and the M.S. and Ph.D. degrees from Princeton University, Princeton, NJ, in 2000, all in electrical engineering. From 2000 to 2005, he was an Assistant Professor with the Computer Science and Engineering Department at Washington University in St. Louis. In 2005, he joined Saint Louis University as an Assistant Professor with the Department of Mathematics and Computer Science. He is the Director of the MediaBench organization, whose goal is continued development and refinement of benchmark suites for multimedia systems research. His work spans a range of topics including image segmentation and objective segmentation evaluation, content-based image retrieval, media processing design and workload evaluation, reconfigurable computer architecture, and multimedia systems.