ASALTAG : Automatic Image Annotation Through Salient Object Detection and Improved k-Nearest Neighbor Feature Matching

Theresia Hendrawati[2], Duman Care Khrisne[1]

[1] Computer Science Graduate Program
Ganesha University of Education (UNDIKSHA)
Singaraja, Bali, Indonesia
theresiahendrawati@gmail.com

[2] Electrical Engineering Department, Faculty of Engineering
Udayana University (UNUD)
Badung, Bali, Indonesia
duman@unud.ac.id

Abstract—Image databases are becoming very large nowadays, and there is an increasing need for automatic image annotation, for assiting on finding the desired specific image. In this paper, we present a new approach of automatic image annotation using salient object detection and improved k-Nearest Neigbor classifier named ASALTAG. ASALTAG is consist of three major part, the segmentation using Minimum Barirer Salienct Region Segmentation, feature extraction using Block Truncation Algorithm, Gray Level Co-occurrence Matrix and Hu' Moments, the last part is classification using improved k-Nearest Neigbor. As the result we get maximum accuracy of 79.56% with k=5, better than earlier research. It is because the saliency object detection we do before the feature extraction proccess give us more focused object in image to annotate. Normalization of the feature vector and the distance measure that we use in ASALTAG also improve the kNN classifier accuracy for labeling image.

***Index Terms*—ASALTAG, Automatic Image Annotation, k-Nearest Neighbor, Salient Region.**

I. INTRODUCTION

Nowadays, image databases are becoming very large and there is an increasing need for automatic tools to automatically annotate and organize image databases. For example average people upload 350 million photos to Facebook per day [1]. The number of images, have impact on the difficulty of people in finding the desired specific image. Information Retrieval researcher using two approach to retrieve or manage such large quantities of images. One using text-based approach and the other use content-based approach. This two approach have its own advantages and disadvantages. To take advantage of each approach the Automatic Image Annotation (AIA) emerge, to automatically annotate each test image with keywords by training a statistical model on a labeled training set [2], [3].

In this paper we try to automaticaly annotate image using baseline approach porposed in [3], and attempted to improve labeling accuracy from what is done in [4]. We first do image segmentation to obtain the object of the image, to do this we use Minimum Barier Detection (MBD) [5]. The salient region extracted from image will be treated as the object to be labeled. Next step is to extract the feature of the image, we use color feature extraction using Block Truncation Algorithm (BTA) [6], the texture feature we use Gray Level Co-occurrence Matrix (GLCM) [7] and for shape feature we use Hu's Moments [8]. Arguably, one of the simplest annotation schemes is to treat the problem of annotation as that of image-retrieval [3]. One can find image nearest neighbor defined by distance feature measure, from the training set, and assign all the keywords of the nearest image to the input test image, so we decide to made feature vector of these image feature and then we use k-NN to clasify image to find it's label, we named the system as ASALTAG. The main contributions of ASALTAG is to introduce new approach to automatically annotate image, which label come from the most salient region of image or the object.

II. Related Work

A number of approach have been proposed for automatic image annotation and retrieval. Among all the proposed approach or models, nearest neighbor are shown to be most successful. Some models successfully used the nearest neighbor, some of them are [3], [4], [9], [10]. Makadia et al. [3] provided the baseline for image annotation based on the nearest neighbor. In [4] Khrisne and Putra use expanded image color-feature by using Block Truncation Algorithm proposed in [6], and with other image feature, they use K-NN to label image. TagProp [9] allows the integration of metric learning by directly maximizing the log-likelihood of the tag predictions in the training set. In this manner, they can optimally combine a collection of image similarity metrics that cover different aspects of image content, such as local shape descriptors, or global color histograms.

In nearest neighbor approach the image feature vector are play important role, they usually used as the characteristic of the image to be classified. Although many methods use the nearest neighbor approach, but the image features are usually obtained from the whole picture as in [3], [4], [6], [9], [10]. Study on image salient region [5], [11] says people pay more attention to the image area that is very contrast with the surrounding environment (salient). If features extracted directly from whole image, then the image feature will contain a lot of information with less meaning, because humans tend to pay attention only to the dominant object in the image. So image salient region segmentation is needed for more accurate label of what people see at the image.

III. Proposed Approach

The proposed work for ASALTAG in this paper is consist of three major part. The segmentation (A), Feature Extraction (B, C, D) and Classification (E).

A. Salient Region Segmentattion

One of the ASALTAG novelty is the image salient region segmentation, before doing feature extraction, using Minimum Barrier Distance (MBD) Transform. In this paper the image features are not obtained from the whole picture, but from part of the image. The part of image is the most salient region or the one that most viewer see as the image's object. For that we use a highly efficient, yet powerful, salient object detection method proposed in [5]. To make saliency map [5] are using Image Boundary Connectivity without using super-pixel representation, for better speed performance. They use MBD to measure image boundary connectivity, and Fast-MBD for better result. We adopt the algoritmh of their system.

Given an input image:

1. For each channel in the Lab color space:
    a. Apply Minimum Barrier Distance transform to compute MBD maps, which measure image boundary connectivity of each pixel
2. Average the MBD maps of the color channels
3. Apply postprocessing to improve the saliency map quality for object segmentation
4. Optionally, we can further enhance the saliency map by leveraging the Backgroundness cue at a moderately increased cost. Backgroundness assumes image boundary regions are mostly background.

After getting saliency map, we make a saliency binary map by thresholding the saliency map. By getting binary map, we can get the results of the image segmentation by multiplying the binary map with the original image.

Apart from the segmentation, the feature extraction method play important role in ASALTAG. ASALTAG are not using a local feature or descriptor because as sugested above, we try to simplify the problem of annotation as of image-retreival in [12]. Global feature are known to be perform better for this task [13]. Whe choose color, texture and shape feature for getting numerical representation of segmented image using method as followed B to D feature extraction part.

B. Block Truncation Algorithm (BTA)

The BTA is an algorithm to extract the color features from an image. Image is divided by the color components R, G and B, the mean of each color component becomes the value for separating the color components into two H and L. Where H for pixels in an image that has a value higher than the average pixel in a color component and L for pixels in an image that has a value lower than the average pixel value in a color component. Thus the color of an image forms 6 groups of RH, RL, GH, GL, BH and BL. The moments of this group are the color features of the BTA. Stricker and Orengo [14], use three central moments of an image's color distribution Mean, Standard deviation and Skewness. In (1), (2) and (3), pk_{ij} is the value of the *k*-th color component of the ij-image pixel and *P* is the height of the image, and *Q* is the width of the image. In ASALTAG whe only use two moment, they are Mean (E_k) and Standart Deviation (SD_k), because we don't want the shape of color distribution, the other feature such texture and shape will replace it.

Moment 1 - Mean :

$$E_k = \frac{1}{PQ}\sum_{i=1}^{P}\sum_{j=1}^{Q} p_{ij}^{k} \quad (1)$$

Moment 2 - Standart Deviation

$$SD_k = \sqrt{\frac{1}{PQ}\sum_{i=1}^{P}\sum_{j=1}^{Q}\left(p_{ij}^{k} - E_k\right)^2} \quad (2)$$

After the extraction process we will get 2 × 6 = 12 color feature.

C. Gray Level Co-occurrence Matrix (GLCM)

GLCM is an image-texture feature extraction technique initiated by [7], it has been utilized as the main tool in image texture analysis in many research paper according to [15], so does in ASALTAG. Haralick suggested statistics

equations that can be calculated from the co-occurrence matrix and be used in describing the image texture. To create a co-occurrence matrix, we need a matrix consisting of reference pixels and neighboring pixels, with distance *d* and angle ɵ, for angel x-axis is the reference for 0° and increasing every 45° counter-clockwise. Fig. 1 show the configuration of working matrix.

		neighbor			
		1	2	…	N
reference	1	1,1	1,2	…	1,N
	2	2,1	2,2	…	2,N
	…	…	…	…	…
	N	N,1	N,2	...	N,N

Fig. 1. The configuration of working matrix

In order to estimate the similarity between different gray level co-occurrence matrices, Haralick [7] proposed 14 statistical features extracted from them. To reduce the computational complexity, only some of these features were selected. The description of 4 most relevant features that are widely used in literature [16], they are Energy (*Er*), Entropy (*En*), Contrast (*Ct*) and Homogeneity (Hg). Equations equations (4) through (7), show us how the texture feature is extracted from the co-occurrence matrix P, with i and j representing the column and row of the matrix P element's and M and N represent number of matrix's coloumn and row.

Texture Feature 1 - Energy :

$$Er = \sum_{i=0}^{M}\sum_{j=0}^{N}(p_{i,j})^2 \quad (4)$$

Texture Feature 2 - Entropy :

$$En = -\sum_{i=0}^{M}\sum_{j=0}^{N} p_{i,j}\log(p_{i,j}) \quad (5)$$

Texture Feature 3 - Contrast :

$$Ct = \sum_{i=0}^{M}\sum_{j=0}^{N}(i-j)^2(p_{i,j}) \quad (6)$$

Texture Feature 4 - Homogeneity :

$$Hg = \sum_{i=0}^{M}\sum_{j=0}^{N}\frac{p_{i,j}}{1+(i-j)^2} \quad (7)$$

In ASALTAG the feature of co-occurrence matrix P is obtained using distance d = 1 to 3 and angle θ of 0°, 45°, 90° and 135°, after feature extraction we get 3 (distance) × 4 (angle) × 4 (feature) = 48 texture feature.

D. Hu's Momment

The moment invariants were first introduced by Hu [8]. Hu moments algorithm is chosen to extract image's shape features since the generated features are rotation, scale and translation invarian. Hu defined seven values, calculated by normalizing central moments completed order three that are invariant to object scale, position, and orientation. In terms of the central moments, the Hu's seven moments are given as shown in (8).

$$\begin{aligned}
\phi_1 &= \mu_{20}+\mu_{02}\\
\phi_2 &= (\mu_{20}-\mu_{02})^2+4\mu_{11}^2\\
\phi_3 &= (\mu_{30}-3\mu_{12})^2+(3\mu_{21}-\mu_{03})^2\\
\phi_4 &= (\mu_{30}+\mu_{12})^2+(\mu_{21}+\mu_{03})^2\\
\phi_5 &= (\mu_{30}-3\mu_{12})(\mu_{21}-\mu_{03})^2[(\mu_{30}+\mu_{12})^2-3(\mu_{21}+\mu_{03})^2]+3(\mu_{21}-\mu_{03})(\mu_{21}+\mu_{03})[3(\mu_{30}+\mu_{12})^2-(3\mu_{21}+\mu_{03})^2]\\
\phi_6 &= (\mu_{20}-\mu_{02})[(\mu_{30}+3\mu_{12})^2-3(\mu_{21}+\mu_{03})^2]-4\mu_{11}\mu_{30}+\mu_{12}(\mu_{21}+\mu_{03})\\
\phi_7 &= 3(\mu_{21}-\mu_{03})+(\mu_{30}+\mu_{12})[(\mu_{30}+\mu_{12})^2-3(\mu_{21}+\mu_{03})^2]-(\mu_{30}-3\mu_{12})+(\mu_{21}-\mu_{03})[3(\mu_{30}+\mu_{12})^2-(\mu_{21}+\mu_{03})^2]
\end{aligned} \quad (8)$$

With Hu's seven moment, ASALTAG get 7 addition feature to be the image characteristics. Now the image's salient segmentation have a 12 + 48 + 7 = 67 feature as deskriptor.

E. Improved K-Nearest Neighbor

K-nearest-neighbor (kNN) classification is one of the most fundamental and simple classification methods and should be one of the first choices for a classification study when there is little or no prior knowledge about the distribution of the data. K-nearest-neighbor classification was developed from the need to perform discriminant analysis when reliable parametric estimates of probability densities are unknown or difficult to determine [17]. The kNN algorithm as in [4] are:

1. Start.
2. Input : Training data, labels for training data, k value, test data.
3. Calculate distance of test data to each training data.
4. Select the training data k which is closest to the test data.
5. Check the label of the k-nearest training data.
6. Determine the label of test data with the most frequent label.
7. Stop.

Differ from [4] that using euclidean-distance for calculating distance between two feature vector, in ASALTAG we decided to improve the distance algorithm. It is because the number of features that we use, the distance is considered as a multi-dimensional distance. The feature vector also have to be normalized first before we calculate the similarity or distance between two feature vector. The distance we use is calculated by (9) proposed by [18].

$$\|A\|_F = \left[\sum_{i,j} abs(a_{i,j})^2\right]^{1/2} \quad (9)$$

IV. RESULT

To implement ASALTAG we use python 2.7 programming language. For image dataset we use the same image dataset with [4], [19], [20] it is Corel Image Dataset. From the dataset we filter the image category and create 11 category with label in Bahasa Indonesia. It is because of the ASALTAG approach, we only take the image from objek-type image in dataset. Fig. 2. Show us when the feature is extracted from single image. As a lot of image consist in one category, we use batch processing for every image on one category, before that we have to set the label for every image in one category, as seen in Fig. 3.

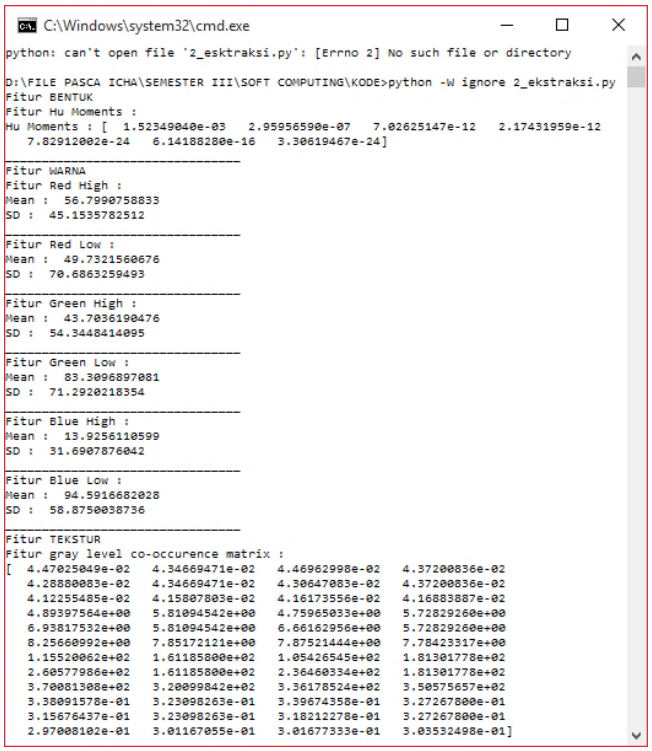

Fig. 2. Feature extraction proccess for single image

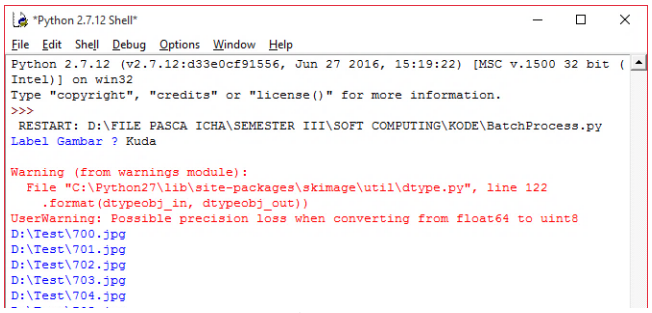

Fig. 3. Batch proccessing feature extraction image category

In total we extract 1925 image for its feature. For testing purpose we use random sampling for every category. with proportion of 67% as training data and 33% as test data for each category. The category and number of image data in dataset can be seen on Table 1.

TABLE I
IMAGE CATEGORY AND DATASET

No.	Label for Category	Number of Image in Dataset
1	*Akordion* (Accordion)	49
2	*Pesawat Terbang* (Airplane)	57
3	*Wajah Manusia* (Human Faces)	435
4	*Bunga Mawar* (Roses)	100
5	*Dinosaurus* (Dinosaur)	97
6	*Bus* (Bus)	100
7	*Kucing* (Cat)	100
8	*Fitness* (Fitnes)	200
9	*Mobil* (Car)	410
10	*Pintu* (Door)	277
11	*Kuda* (Horses)	100

We do about 12 times, kNN classification with random test set, to get the mean accuracy of our system. Fig. 4 show us the process of kNN classification process. We change the value of *k* with 3, 5, 7 and do classification test on data test four times for each *k* value configuration of kNN. The result of this test can be seen on Table 2.

TABLE II
ASALTAG *k*NN LABELING ACCURACY

No.	Training Set	Test Set	*k*-value	Accuracy	Mean Acc for *k* value
1	1286	638	3	78.84%	77.70%
2	1280	644	3	78.10%	
3	1265	659	3	76.32%	
4	1265	659	3	77.54%	
5	1294	630	5	78.73%	**78.17%**
6	1283	641	5	**79.56%**	
7	1282	642	5	75.70%	
8	1299	625	5	78.72%	
9	1260	664	7	78.46%	78.15%
10	1283	641	7	78.62%	
11	1263	661	7	77.15%	
12	1318	606	7	78.38%	
Mean Accuracy				**78.01%**	

Our proposed method in ASALTAG is showing a better accuracy, than what they do in [4], they have maximium-accuracy of 73.26% compared to our maximum-accuracy with 79.56%. however with different k-value our system get average Accuracy of 78.01%. For the k-value, we found that the accuracy is better with $k = 5$.

To get better understanding of our saliency segmentation approach with our ASALTAG expanded kNN, compared to kNN without normalization and using euclidean distance (proposed in [4] and we call it normal kNN) as their similarity check, we do kNN test classification for random test set from same dataset and same scenario, to get the mean accuracy of normal kNN. The result of this test can be seen on Table 3.

From Tabel 2 and Table 3, we found out that ASALTAG kNN accuracy is better that the accuracy of normal kNN. Our way to change kNN distance measure and the

normalisation of the dataset, give us an improvement of kNN way to classify image on dataset. We think the saliency segmentation approach give us more accuracy, as we can see in Table 3. we get better accuracy result than overall accuracy than other earlier research result.

TABLE III
*NORMAL K*NN LABELING ACCURACY

No.	Training Set	Test Set	*k*-value	Accuracy	Mean Acc for *k* value
1	1304	620	3	**69.02%**	64.65%
2	1302	622	3	62.54%	
3	1266	658	3	67.47%	
4	1300	624	3	63.94%	
5	1283	641	5	65.05%	65.75%
6	1246	678	5	67.84%	
7	1246	678	5	62.83%	
8	1285	639	5	67.29%	
9	1313	611	7	67.26%	**67.54%**
10	1291	633	7	66.35%	
11	1289	635	7	68.66%	
12	1276	648	7	67.90%	
Mean Accuracy				**66.10%**	

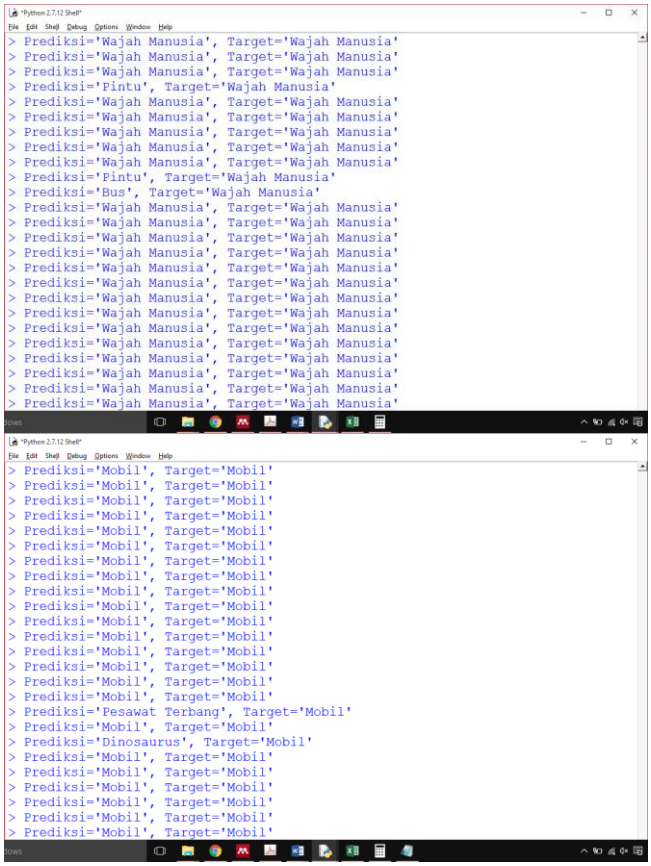

Fig. 3. kNN classification process

V. CONCLUSION

After some test we found that our proposed approach in ASALTAG is getting better accuracy than what the [4] do, using the same classification technique, kNN. It is due to saliency object detection before the feature extraction proccess. Another value we found is, normalization of the feature vector and the distance measure that we use in ASALTAG kNN improve the kNN classifier to get better accuracy for labeling image.

For future research we suggest the use of generatif model to increase accuracy, also we got information of the global and local descriptor combined together can be realy useful for image autolabel.

REFERENCES

[1] V. N. Murthy, E. F. Can, and R. Manmatha, "A Hybrid Model for Automatic Image Annotation," in Proceedings of International Conference on Multimedia Retrieval. ACM, 2014.
[2] M. Ames and M. Naaman, "Why We Tag : Motivations for Annotation in Mobile and Online Media," Proc. SIGCHI Conf. Hum. factors Comput. Syst. ACM, 2007.
[3] A. Makadia, V. Pavlovic, and S. Kumar, "Baselines for Image Annotation," Int. J. Comput. Vis., vol. 1, no. 90, pp. 88–105, 2010.
[4] D. C. Khrisne and D. Putra, "Automatic Image Annotation Menggunakan Block Truncation dan K-Nearest Neighbor" Lontar Komputer, vol. 4, no. 1, pp. 224–230, 2013.
[5] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, and B. Price, "Minimum Barrier Salient Object Detection at 80 FPS," in Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1404–1412.
[6] S. Silakari, M. Motwani, and M. Maheshwari, "Color Image Clustering using Block Truncation Algorithm," Int. J. Comput. Sci. Issues, vol. 4, no. 2, pp. 2–6, 2009.
[7] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Haralick-TexturalFeatures.pdf," IEEE Trans. Syst. Man Cybern., vol. SMC-3, no. 6, pp. 610–621, 1973.
[8] M. K. Hu, "Visual Pattern Recognition by Moment Invariants," IRE Trans. Inf. Theory, vol. 8, no. 2, pp. 179–187, 1962.
[9] M. Guillaumin, T. Mensink, J. Verbeek, C. Schmid, and L. J. Kuntzmann, "TagProp : Discriminative Metric Learning in Nearest Neighbor Models for Image Auto-Annotation," in Computer Vision, 2009 IEEE 12th International Conference, 2009, pp. 309–316.
[10] L. Wu, R. Jin, and A. K. Jain, "Tag Completion for Image Retrieval," IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 3, pp. 716–727, 2013.
[11] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency Optimization from Robust Background Detection," in IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2814–2821.
[12] D. C. Khrisne and M. D. Yusanto, "Content-Based Image Retrieval Menggunakan Metode Block Truncation Algorithm dan Grid Partitioning", S@ CIES vol. 5, no. 2, pp. 79-85, 2015.
[13] C. Singh and P. Sharma, "Performance analysis of various local and global shape descriptors for image retrieval," Multimed. Syst., vol. 19, no. 4, pp. 339–357, 2013.
[14] M. Stricker and M. Orengo, "Similarity of Color Images," in Storage and Retrieval for Image and Video Databases (SPIE), 1995, pp. 381–392.
[15] H. Y. Chai, L. K. Wee, T. T. Swee, S. H. Salleh, and A. K. Ariff, "Gray-Level Co-occurrence Matrix Bone Fracture Detection," Am. J. Appl. Sci., vol. 8, no. 1, p. 26, 2011.
[16] M. Partio, B. Cramariuc, M. Gabbouj, and A. Visa, "Rock Texture Retrieval using Gray Level Co-occurrence Matrix," in Proc. of 5th Nordic Signal Processing Symposium, 2002.
[17] Peterson, Leif E. "K-nearest neighbor." Scholarpedia 4.2 (2009): 1883.
[18] Golub, G. H. and Van Loan, C. F. "Matrix Computations", Baltimore, MD, Johns Hopkins University Press, 1985, pg. 15.
[19] D. Tao, X. Li, and S. J. Maybank, "Negative Samples Analysis in Relevance Feedback," IEEE Transactions on Knowledge and Data Engineering (TKDE), vol. 19, no. 4, pp. 568-580, April 2007
[20] Wang, J. Z., Li, J., and Wiederhold, G., "SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries," IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), vol. 23, no. 9, pp. 947-963, September 2001.