Unsupervised Segmentation Evaluation for Image Annotation

Annette Morales-González[1], Edel García-Reyes[1] and Luis Enrique Sucar[2]
[1]*Advanced Technologies Application Center, Rpto. Siboney, Playa, La Habana, Cuba*
[2]*Instituto Nacional de Astrofísica, Óptica y Electrónica, Puebla, Mexico*

Keywords: Unsupervised Segmentation Evaluation, Automatic Image Annotation, Irregular Graph Pyramid.

Abstract: Unsupervised segmentation evaluation measures are usually validated against human-generated ground-truth. Nevertheless, with the recent growth of image classification methods that use hierarchical segmentation-based representations, it would be desirable to assess the performance of unsupervised segmentation evaluation to select the most suitable levels to perform recognition tasks. Another problem is that unsupervised segmentation evaluation measures use only low-level features, which makes difficult to evaluate how well an object is outlined. In this paper we propose to use four semantic measures, that combined with other state-of-the-art measures improve the evaluation results and also, we validate the results of each unsupervised measure against an image annotation algorithm ground truth, showing that using measures that try to emulate human behaviour is not necessarily what an automatic recognition algorithm may need. We employed the Stanford Background Dataset to validate an image annotation algorithm that includes segmentation evaluation as starting point, and the proposed combination of unsupervised measures showed the best annotation accuracy results.

1 INTRODUCTION

There is a growing tendency to use segmentation-based image representations for object detection and recognition. For instance, the winners of the ILSVRC2013 competition (ImageNet Large Scale Visual Recognition Challenge 2013) (Russakovsky et al., 2014), employed image segmentation in their object detection pipelines. Many other automatic image annotation approaches start from a superpixel representation in order to avoid the complexity of annotation at pixel level (Huang et al., 2011; van de Sande et al., 2011; Arbelaez et al., 2012; Morales-González and García-Reyes, 2013; Zhang and Xie, 2013). In these cases, they use segmentation algorithms to produce an initial partition of the image pixels, but there is always the question of which is a good level of segmentation to start the annotation process for a particular image. If the annotation algorithm uses a hierarchy of image partitions, it is even more important to assess the relevance of each partition to the recognition problem, in order to reduce noise (too over-segmented partitions) and to avoid loosing information (too under-segmented partitions). Nevertheless, the unsupervised segmentation evaluation community has not focused on this particular problem yet.

Unsupervised segmentation evaluation is usually addressed without taking into account the ultimate goal of the task where the segmentation result is to be used. Most approaches compare the results of automatic segmentation against human-annotated ground-truth, or against other automatic segmentation results (Zhang et al., 2008; Csurka et al., 2013). Nevertheless, if segmentation is an initial step in a higher level task, such as image annotation and object recognition, the ground-truth provided by humans may not be adequate in order to obtain good results in the recognition process. It has been suggested by human perception studies (Olson, 2001), that in natural vision the visual input is divided into primitive objects, instead of well defined objects. Therefore, the final output of human perception (an object as a whole) might not be the initial perception cues employed to identify objects or understand scenes. The same applies for automatic recognition, where the final human segmentation (objects as wholes) might not be what a machine needs to perform the classification process. Therefore, evaluating segmentation results in terms of what's more similar to a human segmentation can be misleading in order to know which measure should be used. In this case, it would be desirable that the suitability of a segmentation algorithm would be measured by the success of the end application.

Furthermore, segmentation evaluation is performed using low-level cues of the images, such as edges, color and texture uniformity, inter-region dis-

Morales-González A., García-Reyes E. and Sucar L..
Unsupervised Segmentation Evaluation for Image Annotation.
DOI: 10.5220/0005314201480155
In *Proceedings of the 10th International Conference on Computer Vision Theory and Applications* (VISAPP-2015), pages 148-155
ISBN: 978-989-758-090-1
Copyright © 2015 SCITEPRESS (Science and Technology Publications, Lda.)

parities, etc. (Zhang et al., 2008; Dogra et al., 2012; Khan and Bhuiyan, 2014), but in reality it's very difficult to assess automatically if an image is well segmented or not if there is no knowledge of the semantic entities that are represented there.

In this work, we aim at combining unsupervised segmentation evaluation with an automatic image annotation algorithm, in order to find which are the best image partitions to start the recognition process. The annotation algorithm is based on a segmentation hierarchy, where annotation and segmentation are refined in several iterations, taking advantage of each other information to improve the final recognition result. We propose two contributions: (1) to incorporate semantic information in the unsupervised segmentation evaluation process and (2) to measure the accuracy of the segmentation evaluation in terms of recognition accuracy, which is our ultimate goal. The results obtained in the Stanford Background Dataset (Gould et al., 2009) support these two contributions, showing that the semantic features play an important role in the evaluation process and also, that the behaviour of the segmentation evaluation measures is different when their results are used as input of an image annotation task.

The remaining of this paper includes an analysis of previous works related to unsupervised segmentation evaluation and hierarchical image classification in Section 2. Section 3 describes the hierarchical annotation approach chosen to test the output of segmentation evaluation algorithms. The segmentation evaluation measures used as features in a classification approach to determine good and bad segmentations are presented in Section 4, as well as new measures to take into account semantic information of the classes previously predicted in the image. Section 5 shows the experimental results along with several analysis related to the presented measures which lead to the final conclusions of the paper.

2 RELATED WORK

According to what have been explained in the Introduction, segmentation-based image classification and unsupervised segmentation evaluation are two research fields with a huge disconnection, even though they can be related in some interesting points, in order to improve the results of the former. In this Section, we will analyse several relevant works on both fields to illustrate the necessity of combining them and redirecting the final goal of segmentation evaluation.

2.1 Hierarchical Image Classification Approaches

Many methods for image annotation work at pixel level, but they face several drawbacks, such as the complexity of classifying every pixel of an image and the limited amount of information that a single pixel and its neighborhood may contain (Russell et al., 2014).

Super-pixel based methods have appeared recently, starting with an initial segmentation of the image based on low-level cues and then, performing the annotation over these segments. The advantages over the pixel-based representation can be noted in the possibility of computing better region-based features, and the classification can be performed over a reduced set of entities. Nevertheless, this representation comes with the problem of finding a good segmentation of the image where the annotation process can obtain good results.

Hierarchical models can be used to incorporate local and global image cues. Hierarchical segmentations, in general, are formed by a stack of image partitions (or levels) where each one is built by merging regions from the level below. Therefore, lower levels are over-segmented and higher levels are under-segmented. In (Arbelaez et al., 2012) and (Zhang and Xie, 2013), a hierarchy of segmentations is used for image annotation but no level selection is performed, i.e. they use the entire hierarchy for generating candidate localizations of objects, producing more than 1300 candidates per image. The proposal of (van de Sande et al., 2011) does not select levels either, it starts from the over-segmentation given by a segmentation algorithm and uses the whole hierarchy that's built over that for object detection. The work presented in (Zankl et al., 2012) also deals with the labeling of a segmentation hierarchy, but in this case, they use human input to improve the final image annotation. No automatic segmentation level selection is performed, although the interaction with humans in the labeling process makes it pointless to some extent.

As can be seen, all these methods work with a hierarchy of segmentations, but none of them make an evaluation of the suitability of the segmentation levels employed. Including levels too over-segmented or under-segmented in the recognition process may incorporate noisy partitions and definitely will increase the overall processing cost.

The problem of finding relevant levels in a hierarchy in order to perform object recognition was addressed by (Morales-González and García-Reyes, 2013), by selecting levels that better preserve the edges present in an edge mask of the image (com-

puted using the Canny edge detector), thus inheriting the problems of automatic edge detection, and making the assumption that this edge mask would preserve the actual object boundaries.

2.2 Unsupervised Segmentation Evaluation Methods

In (Zhang et al., 2008) a thorough comparison and analysis of unsupervised segmentation evaluation methods is presented. They performed four different experiments, from which, the closest to our goals is the second one. In this experiment they compared two image partitions that were segmented with the same segmentation method (with different parameters), and the evaluation measures should decide which partition is the best. The measures with better performance in this test were Q, Zeb, and F_{RC} in that order. Q measures the average squared color error of the segments, using penalization terms to decrease the bias towards both over-segmentation and under-segmentation. Zeb uses the internal and external contrast of the regions, measured in the neighborhood of each pixel, to perform the evaluation while F_{RC} takes into account intra-region homogeneity and inter-region disparity.

The $MSET$ evaluation method, reviewed in (Zhang et al., 2008) also, proposed to combine a limited set of measures into a classifier, which outperformed the individual results of these measures for the second experiment. The set of measures employed was composed of E, F, Q, and V_{CP}. E uses region entropy to measure intra-region uniformity and a layout entropy to indicate which pixels belong to which regions. F employs the average color error of the regions, similar to the Q measure mentioned before. V_{CP} uses intra-object measures (e.g. shape regularity, spatial uniformity, etc.), inter-object measures (such as contrast) and each object is weighted by how much attention it received by a human evaluator.

More measures have been introduced recently. In (Morales-González and García-Reyes, 2013) they proposed two measures B_G and B_B that were combined as a weighted sum in order to select the best levels in a hierarchy of segmentations. They are only based on the edges of each partition and how well they match the edges in an edge mask (computed using the Canny edge detector) of the original image. In (Khan and Bhuiyan, 2014) they propose the weighted self-entropy for region homogeneity and the weighted mutual entropy for evaluating region disparity.

In the work presented by (Song et al., 2010) they proposed a method for filtering levels of segmentation in a hierarchy. Levels in the hierarchy are Region Adjacency Graphs (RAGs) and they compute the complexity of the graph on each level using Laplacian graph energy in order to keep those levels whose complexity is smaller than either of the neighboring levels.

All the aforementioned measures work with low-level features (color, texture, edges), but in practice, it is too hard to perform an accurate assessment of the outlining of objects or entities in an image using this information alone. Some kind of higher semantic knowledge should be included in order to know how well the segmentation was performed. Besides, all these measures have been evaluated against human-generated ground truth. The differences of previous approaches and ours are that we are proposing to include segmentation evaluation into an image annotation process in a way that the segmentation evaluation can make use of semantic features coming from predicted classes. Also, the segmentation evaluation measures will be evaluated according to their suitability to predict good levels for the recognition process instead of how well they fit a human-generated ground truth.

3 HIERARCHICAL IMAGE ANNOTATION

We use as base annotation system the approach proposed by (Morales-González et al., 2013), named HMRF-PyrSeg. They use as hierarchical representation the irregular graph pyramids, proposed by (Haxhimusa and Kropatsch, 2004). An irregular graph pyramid is a stack of successfully reduced graphs, where each level is a RAG, i.e. vertices represent regions and the adjacency between them is represented by edges. At the base level, each pixel is a vertex and the edges are the 4-connectivity among pixels. Using a series of edge contractions and eliminations, each level is reduced based on the regions internal and external contrast. The result is a segmentation hierarchy that can be traverse top-down and bottom-up and preserves the topological distribution of the regions throughout all its levels.

Using this representation, the algorithm HMRF-PyrSeg works in the following way (Morales-González et al., 2013):

- The whole graph pyramid is built using low-level cues to segment the image.
- Starting from a predefined level, still over-segmented, every vertex is labeled with a class, using a base classifier (BC) that have been trained previously using low-level features of the regions. This BC must have a probabilistic output (ex.

Random Forest, SVM, Naïve Bayes classifier), which will be used as unary potentials to improve the current BC labeling through a Markov Random Field approach (see next step);

- The whole labeling of this level is improved by means of a Hierarchical Markov Random Field (HMRF), by imposing local constrains among neighboring vertices (in the image plane) and parent/child vertices (in the hierarchy structure);
- A new criterion for edge contraction is used to create a new level of segmentation. This time, the class assigned to each vertex by the HMRF, the probabilities given by the BC to it and the distribution of edges in each partition are combined in order to select which are the vertices that should be joined in the new level. In this way, semantic information coming from the annotation process is combined with low-level information to build more meaningful segmentation levels.
- Once the new level is created, the whole annotation process with the BC and the HMRF is performed again, and this is repeated in several iterations trying to find better image segmentations that yield ultimately to a better recognition result.

Although this approach showed an improvement with respect to other proposals that do not combine the results of annotation and segmentation, problems still arise due to the selection of the first level to begin the whole process. If a level too over-segmented is selected, it can introduce noise in the classification process and this noise will be propagated through all the levels due to the hierarchical information used to classify each level. On the other hand, if an under-segmented level is selected, the boundaries of the objects in the image will be lost and the classification result will suffer as well. According to (Morales-González et al., 2013) they used a fixed level to start this process for all images, disregarding the the nature of each independent image, and displaying the aforementioned problem. An example of this can be seen in Figure 1.

Although the overall result is better, for particular images the results are quite bad. This issue can be addressed by selecting, for each image, the most appropriate level to begin the annotation process. Nevertheless, in the reviewed literature related to segmentation evaluation, all works compared the results of the measures with the human evaluation, which, in this case, doesn't necessarily coincide with what's best for an automatic recognition process.

4 SEGMENTATION EVALUATION IN A HIERARCHY

Since there are many evaluation methods, they measure different aspects of the image partition and they are combined in many different ways (Zhang et al., 2008), we chose the option of using several unsupervised evaluation measures and combine their output values with a classifier. Therefore, a segmentation evaluation classifier (SEC) will be the one who finds out which are the most relevant aspects to be measured and how they should be combined.

4.1 Training Information

In order to obtain training information for our SEC, we use the training set employed for the image annotation process. This training set contains images and their respective irregular pyramid representation (i.e. a hierarchy of segmentations per image). In the HMRF-PyrSeg algorithm, after performing the initial classification of all the regions using the BC, we can know which are the levels, for each image, that obtained better accuracy results when compared with the ground-truth of image annotation. Since our ultimate goal is to improve the annotation results, it sounds natural that the creation of new levels in the HMRF-PyrSeg approach should start from the levels that obtained better accuracy results with the BC. That's why we decided, for each image, to label the n levels that obtained better accuracy with the BC as "good levels", and the rest as "bad levels". We will train a binary classifier with these two labels. We compute all the unsupervised measures for each partition of each hierarchy and we provide these features with their corresponding labels to train the SEC.

4.2 Segmentation Features

We chose several evaluators employed in the literature that measure low-level information of each partition to serve as features that characterize each image segmentation. They can be seen in Table 1.

Since in the HMRF-PyrSeg algorithm we can count with the first classification of each region with a BC, we propose to use semantic information related to this classification in order to add some kind of higher-level information in the evaluation process. These semantic features are 5, 6, 7 and 8 from Table 1.

With the BC, it is possible to have an initial prediction of each region's class. Using this information, our proposals H_r^c (Equation 1) and H_{mr}^c (Equation 2) are the same of H_r and H_{mr} respectively (referred as $H_G(i)$ and $H_G(jk)$ in (Khan and Bhuiyan, 2014)), but

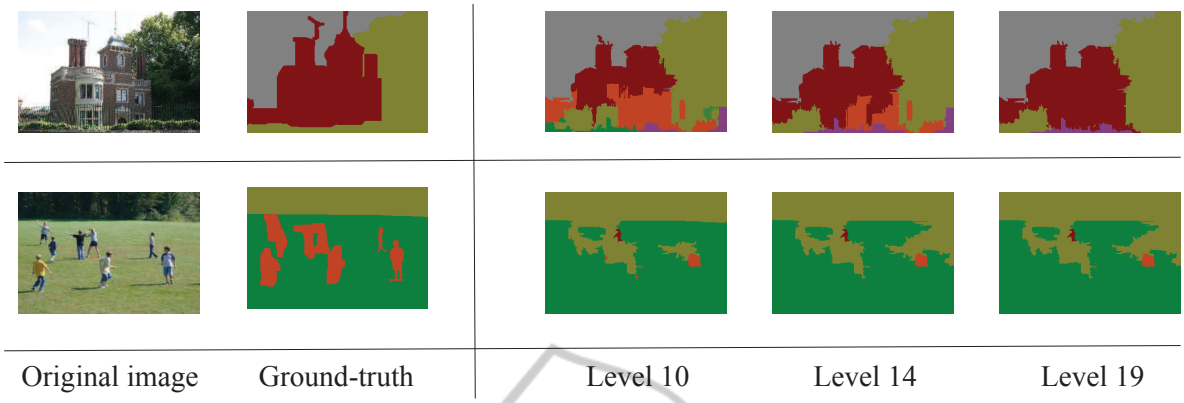

Figure 1: Example image segmentation/annotation results using HMRF-PyrSeg. Level 10 of the pyramid was fixed as starting level for both images. As can be seen, this was a good choice for the first row, where the final result of segmentation/annotation is adequate, but for the second row, level 10 had already lost many meaningful edges.

Table 1: Segmentation evaluation measures used as features for the SEC.

No.	Alias	Description
1	N_r	Number of regions
2	S_r	Average size of regions
3	H_r	Average region's self entropy
4	H_{mr}	Average inter-region's mutual entropy
5	H_r^c	Average class's self entropy
6	H_{mr}^c	Average inter-class's mutual entropy
7	NP_p	Number of pixels with high probability values
8	NP_r	Number of regions with high probability values
9	Zeb	Intra-region and inter-region contrast
10	Q	Squared color error
11	E	Entropy of regions and layout entropy
12	F	Squared color error
13	B_G	Measure of good edges against Canny edge mask
14	B_B	Measure of wrong edges against Canny edge mask
15	P_r	Average perimeter of regions
16	B_n	Number of edge pixels against number of Canny edge pixels

taking into account the whole class area instead of the segmented region. In these equations, G is defined as a feature that describe the pixels (ex. pixel intensity) and $G_{c_i}^{(g)}$ is the set of all possible values of feature G in the area where class c_i was annotated. $N_{c_i}(t)$ is the number of pixels with value t in the c_i class region and M_{c_i} is the total number of pixels in this region. Similarly, Equation 2 uses the same information but for pairwise class analysis, changing the class region being analyzed according to the subscripts in each case. Subscripts c_i, c_j indicates that the region is the union of all the pixels from classes c_i and c_j annotated in the image.

$$H_r^c(c_i) = -\sum_{t \in G_{c_i}^{(g)}} \frac{N_{c_i}(t)}{M_{c_i}} \log \frac{N_{c_i}(t)}{M_{c_i}} \quad (1)$$

$$H_{mr}^c(c_i, c_j) = -\sum_{t \in G_{c_i,c_j}^{(g)}} \frac{N_{c_i,c_j}(t)}{M_{c_i,c_j}} \log \frac{N_{c_i,c_j}(t)}{M_{c_i,c_j}} \quad (2)$$

This means that the entropy is computed in the whole area where the class was detected, disregarding the individual regions that compose that area. In this case, we are measuring the degree of homogeneity in the class area detected (H_r^c) and the disparity among two different classes detected (H_{mr}^c).

The other semantic features related to classification that we are proposing to use are NP_p and NP_r. They employ the probability output of the base classifier. For the case of NP_p we are measuring the amount of pixels that obtained high probability values, normalized by the total number of pixels in the image. NP_r does the same, but instead of taking pixels, they count the amount of regions with high probability values, normalized by the number of regions.

The output of the measures presented in Table 1 can be concatenated into a vector to perform the classification of each image segmentation into "good level" or "bad level". It would be desirable to employ a classifier with probabilistic output, making possible to rank all the scores assigned to the segmentation levels of an image in order to select the best ones.

5 EXPERIMENTS

The experiments were performed in the Stanford Background Dataset (Gould et al., 2009), designed for testing methods developed for geometric and semantic scene understanding. It contains 715 images which

Figure 2: Example images taken from the Stanford Background Dataset. First column shows the original images while second column shows their respective annotation ground-truth, where each color represents one of the 8 semantic labels present in this dataset.

are split in two subsets of 542 and 143 for training and testing respectively. These subsets are randomly generated and the results are averaged. The 8 semantic labels annotated at pixel level are sky, tree, road, grass, water, building, mountain, or foreground object. Two example images from this dataset can be seen in Figure 2 to the left, and their respective annotation ground truth can be seen to the right.

Our objective in these experiments is to find out the influence of choosing each segmentation evaluation measure to select the starting levels of the annotation process. Therefore, the ground truth to evaluate the performance of the measures is given by the accuracy obtained in the annotation process. Although the evaluation results will certainly depend on the selected ground-truth with respect to annotation, this is a common weak point for all annotation/segmentation tasks that are based on a subjective ground-truth created by humans.

The irregular pyramids built for these images usually have around 20 levels, therefore, in order to avoid severe over-segmentation and under-segmentation, we decided to remove some lower and higher levels. In the present case we are analyzing levels from 6 to 16 of each pyramid.

In order to select which is the best combination of measures to use in the SEC, we employed a wrapper feature selection approach (Yang et al., 2013) evaluating different feature subsets with a predictive model. The advantages of using such approach has been stated in (Yang et al., 2013). We exhaustively inspected all possible combinations, skipping combinations of 2 and 3 features only. Since there are few features, the combinatorial explosion is not too high and this can be done in few hours. In this case, the ground truth is the accuracy obtained with the base classifier for the test images. We will consider a good

Table 2: Results of the best measure combinations. First column shows the features employed in each combination, according to the numbering presented in Table 1. Second column shows the accuracy of selecting the one level with highest accuracy of the BC. Third column shows the accuracy of selecting a level among the best three with highest accuracy of the BC.

Feature combination	1 level (%)	3 levels (%)
1, 2, **6**, **7**, 9, 10, 14	**23.78**	**68.53**
1, **6**, **7**, 8, 14, 15, 16	22.38	68.53
1, **6**, **7**, 9, 11, 14, 15	20.28	68.53
1, **6**, **7**, 8, 15, 16	20.28	67.83
1, 4, **6**, 9, 10, 12, 16	18.18	65.73

level selection if the evaluation measure chooses as best level one among the best three with highest accuracy by the BC. If the selected level is not one of the three with highest accuracy, the selection will be considered wrong. Since there were several combinations that achieved the same level selection accuracy, we also measured the accuracy of selecting the one level that has the best BC accuracy. This can be seen in Table 2. The last row of the table is the best combination without taking into account features 7 and 8 from Table 1, since these probabilities may not be available in many approaches. These results were obtained with Random Forest as SEC.

It is important to notice that in the first 4 rows of Table 2, features 1, 6 and 7 are always present, which might indicate that their contribution to the combination is very important. Features 6 and 7 are two of the semantic features proposed in Section 4.2. Also, the difference among the results of the first 4 rows and the 5th row (not using the probabilities of the classifier), also points to the importance of using these semantic features in segmentation evaluation.

Once we had the best feature combination for the SEC, we proceeded to evaluate the performance of each evaluation measure in selecting the best segmentation level for each image. The results of this experiment can be seen in Table 3 and the accuracies shown in second column are from selecting as best level one among the three with highest accuracy of the BC. The measures selected for the experiments were H_r and H_{mr} (Khan and Bhuiyan, 2014), *Zeb*, Q, E, F, reviewed in (Zhang et al., 2008), B, which is the combination of B_G and B_D as presented in (Morales-González and García-Reyes, 2013) and the SEC combination proposed in this work.

According to these results, it can be seen that Q and *Zeb*, which were the measures with better results in Experiment 2 of (Zhang et al., 2008), were not the best for this task. In (Zhang et al., 2008) the ground truth was obtained by human evaluators while in the present task the ground truth was obtained according

Table 3: Results of the level selection accuracy of each measure.

Evaluation measure	Level selection accuracy (%)
H_r	43.36
H_{mr}	13.29
Zeb	39.16
Q	39.16
E	44.06
F	44.76
B	50.35
SEC	**68.53**

to an automatic annotation algorithm output. Therefore, using human-generated ground truth may not be the right assessment to what a computational algorithm needs. It is important to notice the huge improvement displayed by the SEC combination, which outperformed the best result in 18 % of accuracy. This is an indicator of the benefits provided by the combination and the use of semantic features.

Additional information regarding these measures can be seen in Table 4. Second column shows the average time to evaluate all the segmentation levels of one image (currently 11 levels, from level 6 to 16 of each pyramid). Another interesting information is how deviated to under or over-segmentation each measure is, with respect to the ground-truth correct levels. We computed the difference between the best levels selected by each measure and the ground-truth levels, and computed the mean and standard deviation of this difference. These values are shown in columns 3 and 4 respectively. A negative mean value indicates that the corresponding measure tends to over-segmentation w.r.t the ground-truth correct levels. Conversely, a positive value indicates that the measure tends to select under-segmented levels. Values closer to zero correspond to measures with outputs closer to the ground-truth. In this sense, it can be seen that most measures, except for H_{mr}, tend to choose levels more over-segmented (at different degrees) than the ground-truth. H_{mr} has a strong bias to under-segmentation while H_r and *F* have the stronger biases to over-segmentation. The SEC combination displays the mean value closest to zero with a slight bias to over-segmentation, and the lowest standard deviation among all the measures.

Regarding the computational cost of computing each individual measure, it can be seen in Table 4 that *Zeb* is the more time-consuming measure, followed by *B* and H_{mr}. The time shown for the SEC combination corresponds to the second combination presented in Table 2. The reason for this is that the first combination employs the *Zeb* measure, which greatly increases the computation time (10.58 seconds). It

Table 4: Additional information of each measure.

Evaluation measure	Time (s)	Mean	Stdev
H_r	0.702	-2.34	3.38
H_{mr}	2.271	4.85	2.74
Zeb	6.296	-0.69	4.73
Q	0.214	-1.73	4.12
E	**0.206**	-0.89	3.44
F	0.215	-2.55	2.84
B	3.952	-1.86	2.72
SEC	4.505	**-0.67**	**2.43**

Table 5: Accuracy of the annotation process when choosing as starting levels the ones selected by each measure.

Evaluation measure.	Annotation accuracy (%)
Base Classifier	73.0
Fixed Level	75.2
H_r	72.34
H_{mr}	63.78
Zeb	71.41
Q	69.75
E	71.63
F	71.91
B	74.72
SEC	**76.86**

is important to notice that the SEC combination employs other measures that also contribute to the total time. Nevertheless, since many of these individual measures work with common information, they can be computed together, reducing the total time with respect to the sum of their individual times. Also, the cost and the accuracy information can be used to find an appropriate trade-off between these two aspects in specific applications.

Using as starting levels the ones selected by each measure, we ran the whole annotation algorithm and the final annotation accuracy in each case can be seen in Table 5. Also, the first two rows show the annotation accuracy of the base classifier and the annotation accuracy of the hierarchical annotation process using a fixed level (level 10 in this case) for starting the annotation process.

In this experiment can be seen that, in most cases, selecting the starting levels with the segmentation evaluation measures deteriorates the final annotation accuracy with respect to the base classifier accuracy and using a fixed level. The only measures that improved the BC results were *B* and the SEC combination, while the fixed level approach was only outperformed by the SEC combination. here is a significant improvement of the SEC combination over the second best evaluation measure (*B*), of 2.14 %.

SCITEPRESS SCIENCE AND TECHNOLOGY PUBLICATIONS

6 CONCLUSIONS

In this paper we addressed two usually unrelated research fields: unsupervised segmentation evaluation and automatic image annotation. Our proposal of including semantic measures in the segmentation evaluation process and the combination of several individual evaluators displayed better results than the most relevant measures found in the literature. We also showed that the measures that evaluate a segmentation more similar to humans, are not the best for selecting partition levels to perform automatic recognition tasks. Therefore, in our opinion, more efforts should be devoted to develop segmentation evaluation measures that work better for automatic image annotation, instead of focusing on the best segmentation output for humans.

The final results of the annotation process showed that selecting "good" levels at the beginning provides better annotation accuracy. As future work, we plan to include saliency maps in the segmentation evaluation process, trying to find partitions that preserve distinctive objects or parts.

ACKNOWLEDGEMENTS

This work was supported in part by CONACYT project 215546.

REFERENCES

Arbelaez, P., Hariharan, B., Gu, C., Gupta, S., Bourdev, L. D., and Malik, J. (2012). Semantic segmentation using regions and parts. In *CVPR*, pages 3378–3385. IEEE.

Csurka, G., Larlus, D., and Perronnin, F. (2013). What is a good evaluation measure for semantic segmentation? In *24th British Machine Vision Conference (BMVC)*, University of Bristol, United Kingdom.

Dogra, D. P., Majumdar, A. K., and Sural, S. (2012). Evaluation of segmentation techniques using region area and boundary matching information. *J. Vis. Comun. Image Represent.*, 23(1):150–160.

Gould, S., Fulton, R., and Koller, D. (2009). Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, pages 1–8. IEEE.

Haxhimusa, Y. and Kropatsch, W. G. (2004). Segmentation graph hierarchies. In *Proceedings of Joint International Workshops on Structural, Syntactic, and Statistical Pattern Recognition S+SSPR 2004*, volume LNCS 3138, pages 343–351. Springer, Berlin Heidelberg, New York.

Huang, Q., Han, M., Wu, B., and Ioffe, S. (2011). A hierarchical conditional random field model for labeling and segmenting images of street scenes. *2013 IEEE Conference on Computer Vision and Pattern Recognition*, 0:1953–1960.

Khan, J. F. and Bhuiyan, S. M. (2014). Weighted entropy for segmentation evaluation. *Optics and Laser Technology*, 57(0):236 – 242. Optical Image Processing.

Morales-González, A. and García-Reyes, E. B. (2013). Simple object recognition based on spatial relations and visual features represented using irregular pyramids. *Multimedia Tools Appl.*, 63(3):875–897.

Morales-González, A., Reyes, E. B. G., and Sucar, L. E. (2013). Improving image segmentation for boosting image annotation with irregular pyramids. In *CIARP (1)*, volume 8258 of *LNCS*, pages 399–406. Springer.

Olson, C. R. (2001). Object-based vision and attention in primates. *Current Opinion in Neurob.*, 11:171–179.

Russakovsky, O., Deng, J., Krause, J., Berg, A., and Li, F. (2014). Results of ILSVRC2013. http://www.image-net.org/challenges/LSVRC/2013/results.php.

Russell, C., Ladicky, L., Kohli, P., and Torr, P. H. S. (2014). Associative hierarchical random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6):1–1.

Song, Y.-Z., Arbelaez, P., Hall, P. M., Li, C., and Balikai, A. (2010). Finding semantic structures in image hierarchies using laplacian graph energy. In *ECCV (4)*, volume 6314 of *LNCS*, pages 694–707. Springer.

van de Sande, K. E. A., Uijlings, J. R. R., Gevers, T., and Smeulders, A. W. M. (2011). Segmentation as selective search for object recognition. In *Proceedings of ICCV '11*, pages 1879–1886. IEEE Computer Society.

Yang, P., Liu, W., Zhou, B. B., Chawla, S., and Zomaya, A. Y. (2013). Ensemble-based wrapper methods for feature selection and class imbalance learning. In *PAKDD (1)*, volume 7818 of *LNCS*, pages 544–555. Springer.

Zankl, G., Haxhimusa, Y., and Ion, A. (2012). Interactive labeling of image segmentation hierarchies. In *DAGM/OAGM Symposium*, volume 7476 of *LNCS*, pages 11–20. Springer.

Zhang, H., Fritts, J. E., and Goldman, S. A. (2008). Image segmentation evaluation: A survey of unsupervised methods. *Comput. Vis. Image Underst.*, 110(2):260–280.

Zhang, S. and Xie, M. (2013). Beyond sliding windows: Object detection based on hierarchical segmentation model. In *International Conference on Communications, Circuits and Systems (ICCCAS)*, pages 263 – 266. IEEE.