LETTER

# Normalized Joint Mutual Information Measure for Ground Truth Based Segmentation Evaluation

Xue BAI[†a)], Yibiao ZHAO[†b)], *Nonmembers*, *and* Siwei LUO[†c)], *Member*

**SUMMARY** Ground truth based image segmentation evaluation paradigm plays an important role in objective evaluation of segmentation algorithms. So far, many evaluation methods in terms of comparing clusterings in machine learning field have been developed. However, most traditional pairwise similarity measures, which only compare a machine generated clustering to a "true" clustering, have their limitations in some cases, e.g. when multiple ground truths are available for the same image. In this letter, we propose utilizing an information theoretic measure, named NJMI (Normalized Joint Mutual Information), to handle the situations which the pairwise measures can not deal with. We illustrate the effectiveness of NJMI for both unsupervised and supervised segmentation evaluation.

***key words:*** *image segmentation evaluation, similarity measure, joint mutual information*

## 1. Introduction

Image segmentation is an indispensable pre-processing step in many vision systems. Many efforts have been devoted to developing more effective segmentation techniques, as well as quantifying the performance of current algorithms. However, due to the ill-defined nature of the segmentation problem, evaluation of segmentation results is still a challenging task. In order to obtain more objective evaluation scores instead of just using subjective judgments, a database of human segmented natural images [1] was established. Therefore, based on the "true" segmentations, ground-truth-based (GT-based) evaluation paradigm is preferred. In this paradigm, most evaluation methods can be either region-based or boundary-based. Here, we focus on the methods used for evaluating region-based segmentation algorithms.

Since region segmentation can be seen as a clustering procedure for image pixels according to the feature vector for each pixel including color and spacial information, it is a natural way to carry out evaluation tasks in terms of clusterings comparison, i.e. compare the machine outputs against the ground-truth segmentations through some measure of similarity. So far, a lot of clustering-comparison measures have been proposed in machine learning domain, and they can be categorized into three classes which are pair-counting based (e.g. Rand Index [2]), set-matching based (e.g. $\mathcal{H}$ criterion [3]), and information theoretic based similarity measures (e.g. Normalized Mutual Information [4] and Variation of Information [5]). Jiang et al. [6] applied various clustering-comparison measures to GT-based segmentation evaluation, based on both range images and intensity images, and the experimental results demonstrate their usefulness and applicability in quantifying the performance of segmentation algorithms.

In practice, it is not always possible to compare only two segmentations, e.g. when multiple ground truths are available for the same image, or the evaluated algorithm could generate more than one segmentation to be matched with a ground truth, so the pairwise similarity measures should be extended to deal with more general situations. In this article, we propose an information theoretic based measure, the Normalized Joint Mutual Information (NJMI), which is an extension to the Normalized Mutual Information (NMI) to handle the general cases mentioned above.

The rest of this article is organized as follows. In Sect. 2, we first describe the Joint Mutual Information (JMI) and its normalized version NJMI. In Sect. 3 and Sect. 4, we present how the NJMI is useful in ground truth based evaluation for both unsupervised and supervised segmentation respectively. Section 5 illustrates how to eliminate the "outlier" ground-truth segmentation from crowdsourced annotations through NJMI. Section 6 gives the conclusion.

## 2. Normalized Joint Mutual Information for Multiple-to-One Clusterings Comparison

In [7], the Joint Mutual Information (JMI) has been proposed to be applied to feature selection. Given a target variable $Y$ and a set of input variables $X = \{X_1, \ldots, X_n\}$, the relevance between one feature $Y$ and other features $X$ can be defined by the JMI:

$$I(X_1, \ldots, X_n; Y) = KL(P(X_1, \ldots, X_n, Y) \| P(X_1, \ldots, X_n)P(Y)) \quad (1)$$

where $KL(\cdot\|\cdot)$ denotes the Kullback-Leibler divergence. So, the JMI can be written as:

$$I(X_1, \ldots, X_n; Y) = \sum_{X_1,\ldots,X_n,Y} P(X_1, \ldots, X_n, Y) \log \frac{P(X_1, \ldots, X_n, Y)}{P(X_1, \ldots, X_n)P(Y)} \quad (2)$$

Therefore, the JMI can be seen as an extension to the mutual information presented in Appendix.

Manuscript received April 9, 2012.

Manuscript revised June 28, 2012.

[†]The authors are with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China.

a) E-mail: bjtuxbai@gmail.com

b) E-mail: ybzhao.lhi@gmail.com

c) E-mail: swluo@bjtu.edu.cn

DOI: 10.1587/transinf.E95.D.2581

Like mutual information, JMI does not have a fixed upper bound. To make the evaluation scores comparable in a fixed range $[0, 1]$, we need a normalized version of JMI. In Appendix A, the normalized mutual information is given by Eq. (A·2). In much the same way, we infer that JMI is bounded by the entropy $H(Y)$ and the joint entropy $H(X_1, \ldots, X_n)$. So, the Normalized Joint Mutual Information (NJMI) is defined as

$$NJMI(X_1, \ldots, X_n; Y) = \frac{I(X_1, \ldots, X_n; Y)}{\sqrt{H(X_1, \ldots, X_n)H(Y)}} \qquad (3)$$

where $H(Y) = -\sum_Y P(Y)\log P(Y)$ and $H(X_1, \ldots, X_n) = -\sum_{X_1,\ldots,X_n} P(X_1, \ldots, X_n)\log P(X_1, \ldots, X_n)$.

## 3. NJMI for Unsupervised Segmentation Evaluation

### 3.1 Multiple Ground Truths

In a hand-labeled segmentations dataset, for the same image, different human subjects always produce different segmented results due to subject prior knowledge, or simply at various granularity levels, leading to more than one ground-truth segmentations for an image, e.g. Berkeley segmentation database [1]. Figure 1 gives an example image and its five manually segmented images.

Therefore, for the task of unsupervised segmentation evaluation with multiple manually labeled images, we propose to use NJMI to measure the similarity between the segmentation generated by an algorithm and a set of ground-truth images. Given an image $I$ including $N$ pixels, if set $X = \{X_1, \ldots, X_n\}$, denotes a set of ground-truth segmentations, variable $Y$ denotes a segmentation compared with $X$, the similarity measure can be calculated by Eq. (3).

### 3.2 Experiment

We first present the performance of NJMI on selecting an appropriate parameter setting for an algorithm. Figure 2 shows seven mean shift [8] segmentations (from oversegmentation to undersegmentation) using different bandwidth parameters. Figure 3 depicts three evaluation scores, Global Consistency Error (GCE), Local Consistency Error (LCE) [1], and NJMI over the segmentations (a)-(g) in Fig. 2. From this plot, we observe that NJMI can correctly reflect that (e) is the best segmentation, (c) and (d) are weakly acceptable, (b) is on the borderline, and (a), (f), and (g) are the worst.

Furthermore, we explore the segmentation performance of three algorithms, mean shift (MS) [8], efficient graph (EG) [9] and normalized cut (NC) [10] through NJMI. For each algorithm, we select the best segmentation result from four parameter settings, and five images covering different types of objects are tested. As shown in Fig. 4, the segmentation result getting the highest value of NJMI by MS in image2 is relatively more consistent with the ground truths, while the segmentation with the lowest NJMI value obtained by MS in image1 hardly depict the objects.

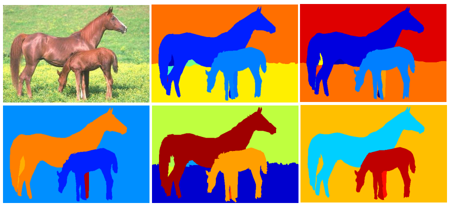


**Fig. 1** An example image and its five ground-truth segmentations.

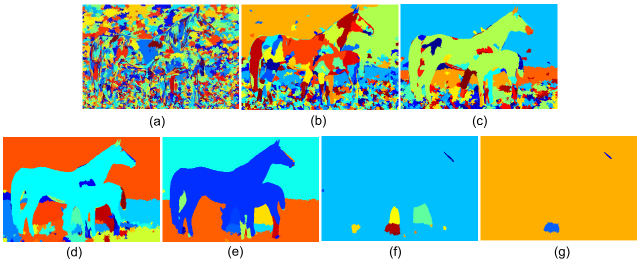


**Fig. 2** Seven mean shift segmentations.

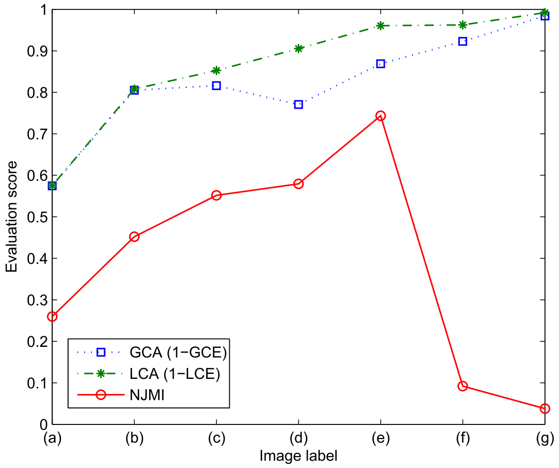


**Fig. 3** Three evaluation scores for different segmentations: Global Consistency Accuracy (GCA), Local Consistency Accuracy (LCA), and NJMI.

## 4. NJMI for Supervised Segmentation Evaluation

### 4.1 Stability Evaluation of Interactive Segmentation Algorithms

As the common automatic segmentation algorithms without high level prior knowledge of interest object always can not obtain satisfied results, the interactive segmentation paradigm incorporating human interaction draws more attention recently. A lot of state-of-the-art interactive segmentation algorithms have been developed in the last decade. Thus, appropriate evaluation for interactive or supervised segmentation performance is indispensable.

Being different from unsupervised segmentation, evaluating the segmentation performance with human interference should also consider the stability of algorithms, as well

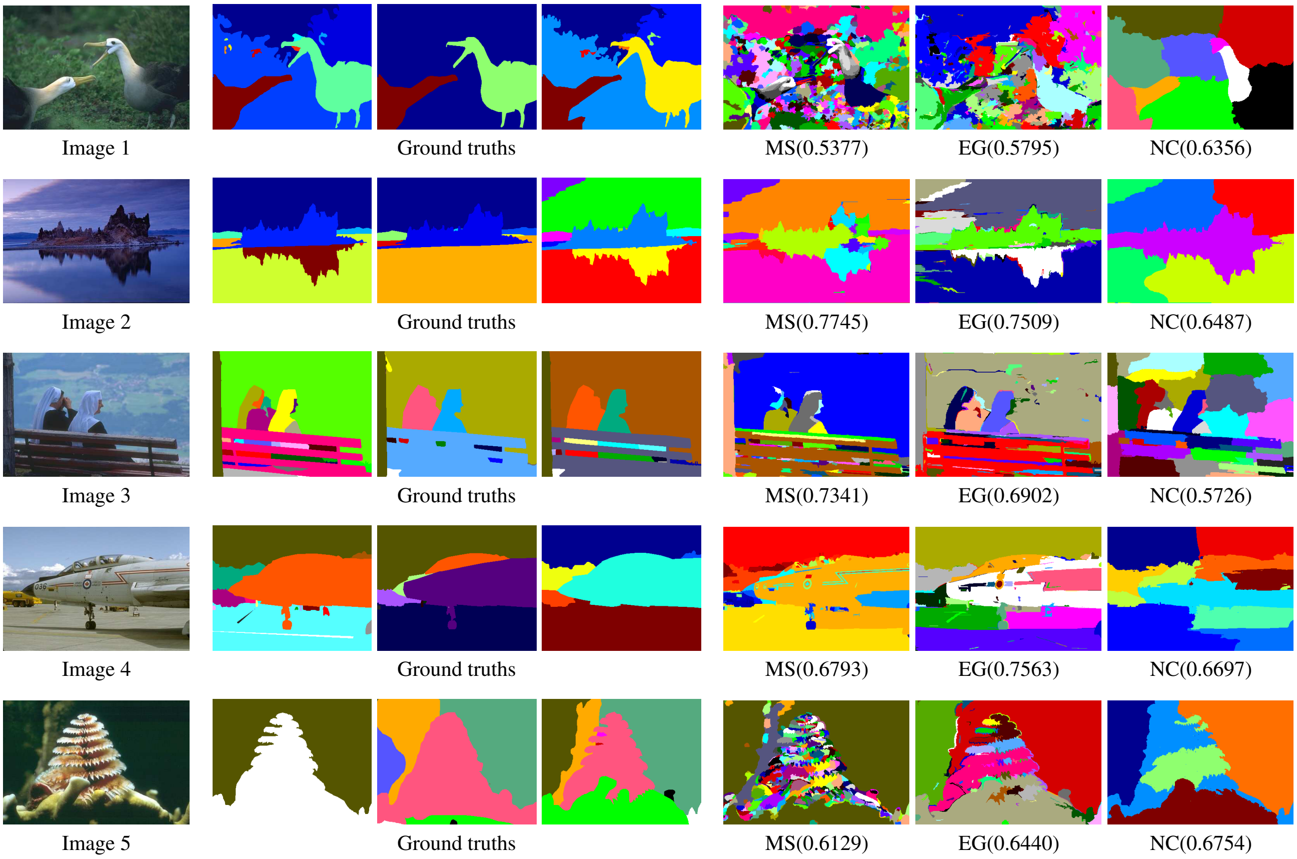


**Fig. 4** Compare three segmentation algorithms by NJMI.

as accuracy. For instance, if users label some key components of an image to indicate the major features of the foreground and background, and expect that the algorithm can predict desired labels to the remaining parts of the image, for an stable algorithm, the more interaction, the better performance is achieved, while the segmentation results generated by unstable algorithms change a lot by minor alteration of interactions. In this context, multiple machine segmentations corresponding to different interactive settings should be involved in quantitatively evaluating the interactive algorithms by comparing with a ground truth image.

### 4.2 Experiment

To objectively simulate interactive process of drawing scribbles, a point-process of human interaction is integrated in segmentation procedure, which draws points on key components of foreground/background. In Fig. 5, the first row on the right four columns shows four levels of point scribbles increasing from left to right, and the next two rows are corresponding interactive segmentation results by two algorithms (random walker and structural IS) [11]. We can see that, structural IS with higher NJMI value is more stable and has better segmentation performance than random walker.

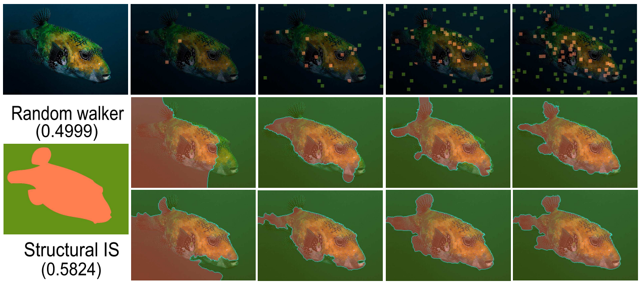
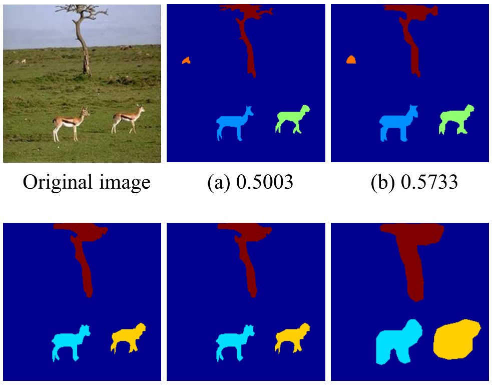


**Fig. 5** Evaluation of interactive segmentation algorithms.

## 5. NJMI for Ground Truth Evaluation

### 5.1 Quality Evaluation of Crowdsourced Annotations

As we see that, to achieve reasonable segmentation evaluation, the quality of ground truth is also a key problem. In computer vision field, with the increasing interest in larger and diverse object data sets, some annotation data sets mainly relying on non-expert, online annotators have been established, such as LabelMe [12]. Due to the heterogeneous and unpredictable capability of minimally trained annotators, it is significant to develop strategies for automatically estimating the quality of these annotations or crowdsourced ground truths. For this task, we can also use NJMI



Original image (a) 0.5003 (b) 0.5733

(c) 0.6400 (d) 0.6400 (e) 0.4520

**Fig. 6** Evaluation of annotations.

to eliminate the "ourlier" annotation from multiple segmentations annotated by different users.

### 5.2 Experiment

As illustrated in Fig. 6, there are five annotations for the given image, and (e) is the "outlier" among them. For each annotation, we compare it with others and obtain the NJMI value. We can observe that, the lowest NJMI value indicates the "outlier" annotation.

## 6. Conclusion

In this letter, we presented the usefulness of an information theoretic based measure, the Normalized Joint Mutual Information (NJMI), in more general cases of ground truth based segmentation evaluation. Through the experiments, we observe that this evaluation index can give reasonable scores for comparing "one with multiple", where the pairwise measures can not be used.

## Acknowledgments

This work is supported by National Nature Science Foundation of China (60975078, 60902058, 61105119) and Beijing Natural Science Foundation (4112047).

### References

[1] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," ICCV, 2001.

[2] W.M. Rand, "Objective criteria for the evaluation clustering methods," J. American Statistical Association, vol.66, no.336, pp.846–850, 1971.

[3] M. Meilă and D. Heckerman, "An experimental comparison of model-based clustering methods," Mach. Learn., vol.42, no.1-2, pp.9–29, 2001.

[4] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," J. Machine Learning Research, vol.3, pp.583–617, 2002.

[5] M. Meilă, "Comparing clusterings by the variation of information," Conf. Learning Theory, pp.173–187, 2003.

[6] X. Jiang, C. Marti, C. Irniger, and H. Bunke, "Distance measures for image segmentation evaluation," EURASIP Journal on Applied Signal Processing, pp.1–10, 2006.

[7] H.H. Yang and J. Moody, "Data visualization and feature selection: New algorithms for nongaussian data," Advance in Neural Information Processing Systems 12, 1999.

[8] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," IEEE Trans. Pattern Anal. Mach. Intell., vol.24, no.5, pp.603–619, 2002.

[9] D.H.P. Felzenszwalb, "Efficient graph-based image segmentation," Int. J. Comput. Vis., vol.59, no.2, pp.167–181, 2004.

[10] J. Shi and J. Malik, "Normalized cuts and image segmentation," IEEE Trans. Pattern Anal. Mach. Intell., vol.22, no.8, pp.888–905, 2000.

[11] Y. Zhao, X. Nie, Y. Duan, Y. Huang, and S. Luo, "A benchmark for interactive image segmentation algorithms," 2011 Workshop on Person-Oriented Vision (POV), pp.33–38, 2011.

[12] K.P. Murphy, C. Russell, A. Torralba, and W.T. Freeman, "Labelme: a database and web-based tool for image annotation," Int. J. Comput. Vis., vol.77, pp.157–173, 2008.

[13] T.M. Cover and J.A. Thomas, Elements of information theory, Wiley, 1991.

## Appendix: Normalized Mutual Information for Binary Clusterings Comparison

In machine learning, Normalized Mutual Information (NMI) is a widely used information criteria for clusterings comparison. It measures the similarity or distance between two clusterings by evaluating the mutual information between them.

Let $D$ be a set of $N$ data points $\{d_1, \ldots, d_N\}$, and $X$, $Y$ are two clusterings for $D$, where $X$ includes $r$ clusters $\{x_1, \ldots, x_r\}$, and $Y$ includes $c$ clusters $\{y_1, \ldots, y_c\}$. If we regard $X$ and $Y$ as two random variables of cluster labels, $P(x_i) = |x_i|/N$, $P(y_j) = |y_j|/N$ are the probabilities of a random data point labeled by $x_i$ in $X$ and labeled by $y_j$ in $Y$ respectively, and $P(x_i, y_j) = |x_i \cap y_j|/N$ represents the joint probability that a data point labeled by $x_i$ in $X$ and $y_j$ in $Y$ simultaneously, then according to information theory [13], the mutual information between random variables $X$ and $Y$ is calculated as

$$\begin{aligned} I(X;Y) &= \sum_{X,Y} P(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)} \\ &= \sum_{i=1}^{r} \sum_{j=1}^{c} P(x_i, y_j) \log \frac{P(x_i, y_j)}{P(x_i)P(y_j)} \qquad \text{(A$\cdot$ 1)} \end{aligned}$$

Further more, [4] proposed a normalized version of the mutual information which has fixed bounds [0, 1]:

$$NMI(X;Y) = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}} \qquad \text{(A$\cdot$ 2)}$$

where $H(X) = -\sum_X P(X) \log P(X) = -\sum_{i=1}^{r} P(x_i) \log P(x_i)$ and $H(Y) = -\sum_Y P(Y) \log P(Y) = -\sum_{j=1}^{c} P(y_j) \log P(y_j)$ are the entropies associated with $X$ and $Y$ respectively.