Adaptive Figure-Ground Classification by Statistical Distance Measure

Yisong Chen, Hong Cui, Antoni Chan

Abstract

1. Introduction

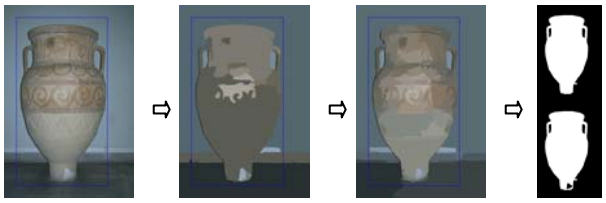
Foreground extraction in still images plays a key role in vision applications [1]. We present a weakly supervised foreground extraction framework that gives promising solutions in a broadly applicable environment. The pipeline is illustrated in Figure 1. Under the assumption that a bounding-box mask is able to provide sufficient statistical information about the background, we treat the task as a figure-ground (f-g) classification on the over-segmented patches generated by the adaptive mean-shift algorithm [2]. We model all the region patches as multivariate normal distributions in a 5D joint color-spatial feature space. Two novel probability distances are defined to measure the similarities and new labels are assigned progressively by statistical distance comparison. Multiple hypotheses are output to add the chance of success. This scheme avoids the trouble of parameter tuning and makes it possible to fully enjoy the favorable characteristics of the mean-shift algorithm in a direct and intuitive manner. It overcomes many drawbacks of state-of-the-art techniques and generates surprisingly good results for challenging images. *The main contribution is a very simple model equipped with two powerful distance measures, which leads to efficient solving procedure and excellent results*.

2. A figure-ground classification framework

We call our algorithm a weakly supervised one because it merely relies on interactive mask assigning and need no other a prior knowledge. Briefly speaking, a mask bounding-box is interactively assigned by the user as [1]. Either side of the box can be defined as the background mask. The complement of the background mask makes the foreground mask. This mask definition flexibly handles different cases of partially-inside foreground. Some example bounding-boxes are illustrated in Figure 1 and 2.

2.1. Patch making: adaptive mean-shift

Defining the segmentation as the grouping of non-overlapping regions instead of pixels has become a popular approach due to its advantages in information transfer and computational efficiency. We choose mean-shift as our super-pixel generator because mean-shift patches are easier to describe statistically. The two bandwidth parameters h_s/h_r are adaptively initialized using the relationship between the bandwidth parameters and the covariance matrix of the multivariate distribution [3].



(a) (b) (c) (d)

Figure 1. Adaptive figure-ground classification solving pipeline
(a) original image & mask; (b) initial ms patches(h_s=*7*,h_r=*6.0*)
(c) adaptive ms patches(h_s=*15*,h_r=*2.7*); (d) D_M and D_K selections

We directly adopt the mean-shift 5D space as our feature space for similarity measure. In other words, we treat the 3D color features and the 2D spatial features identically and do not give any priority to spatially adjacent patches. A feature vector in the 5D feature space is given by

$$f = (L, a, b, x, y) \quad (1)$$

where (x y) are the 2D pixel coordinates and (L a b) are the pixel values in the Lab color space. We model each mean-shift patch p_i as a multivariate normal distribution $N(\mu_i, \Sigma_i)$ in the 5D feature space. The 5D mean vector μ_i and the 5*5 covariance matrix Σ_i are estimated using patch statistics. For accuracy all the patches are eroded with a radius-1 disk structuring element to avoid border effects.

2.2. Similarity measure: statistical distance

After the adaptive mean-shift, we label the patches overlapping with the background mask region as the background priors, and obtain an initial foreground map. The final foreground is obtained by gradually refining the initial foreground map via statistical distance comparison.

The multivariate Gaussian model makes it easy to measure the probability distance between two mean-shift patches. It is well known that there exists a closed-form Kullback-Leibler(KL) divergence between two Gaussians $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$ [4].

$$KL(N_1, N_2) = \frac{1}{2}(\log(|\Sigma_2|/|\Sigma_1|) + Tr(\Sigma_2^{-1}\Sigma_1) + (\mu_1 - \mu_2)^T \Sigma_2^{-1}(\mu_1 - \mu_2)) \quad (2)$$

Equation (2) is not symmetric and thus inconvenient in similarity comparison. To overcome this drawback we suggest the following minimum KL-divergence to measure the statistical distance.

$$D_K(N_1, N_2) = \min(KL(N_1, N_2), KL(N_2, N_1)) \quad (3)$$

Equation (3) is a symmetrized variation of the KL divergence between two Gaussians. It has an intuitive interpretation that the two patches should be grouped together if either of them can be well described by the other.

The computation of the logarithm term of Equation (2) is sometime numerically instable due to unreliable covariance matrixs Σ_1 or Σ_2 caused by singular patches. To remove

such instability we also define a more conservative minimum Mahalanobis distance.

$$D_M(N_1,N_2)=\min((\mu_1-\mu_2)^T\Sigma_2^{-1}(\mu_1-\mu_2),(\mu_2-\mu_1)^T\Sigma_1^{-1}(\mu_2-\mu_1)) \quad (4)$$

Equation (4) can be deemed as a variation of the minimum KL divergence by retaining only the dominant mode comparison term. Both D_M and D_K treat the mutual "belong to" relationship well and the background holes can be reliably identified. Roughly speaking, there is no guarantee one of them is better than the other. But they indeed provide beneficial complements to each other. Therefore, in our approach we take both similarity measures and output multiple hypotheses.

Provided a similarity measure D (either D_M or D_K), we can define the distance from a single patch *p* to a region set *R* by equation (5), and the distance between two region sets *R1* and *R2* by equation (6).

$$D(p,R)=\min_{r\in R}(D(p,r)) \quad (5)$$

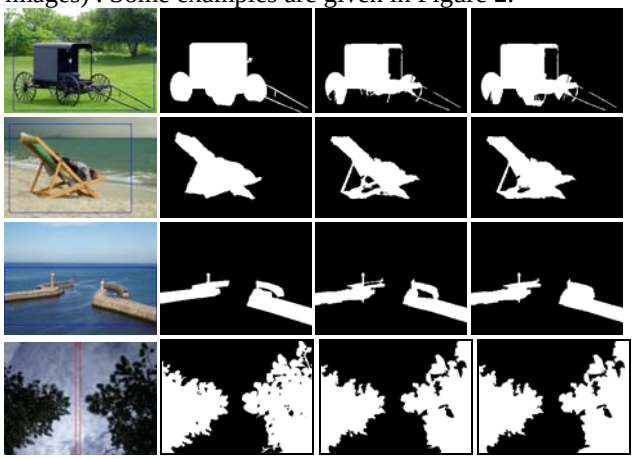
$$D(R_1,R_2)=\min_{r\in R_1}(D(r,R_2)) \quad (6)$$

2.3. Binary classification: gradual labeling

A full classification trial is composed of the following steps. First, the patches sufficiently far from the background priors are labeled as foreground patches. Second, the unlabeled patches are gradually merged into the foreground or the background group by comparing their distances to the existing foreground patches and background priors. Finally, an evaluation score is optimized to select the most promising solutions by maximizing the global distance (5) or (6) between background and foreground patches.

3. Experiments

We test our method on three popular datasets (100 Weizmann 1-obj, 100 Weizmann 2-obj, and 50 grabcut test images) . Some examples are given in Figure 2.



original(with mask) ground truth D_M selection D_K selection

Figure 2. Some Weizmann image set results. The blue boxes enclose foreground mask and the red box encloses the background mask.

The algorithm outputs all candidates selected by both D_M and D_K and leaves the final decision to the user. It is powerful in labeling background holes or multiple connected components, which are sometime even missed in the manual-made truths. Table 1 reports the comparison with the grabcut algorithm [1] in terms of the 95% confidence intervals of the F-measure, *F=2PR/(P+R)*, where *P* and *R* are the precision and recall values.

Table 1. F-measure comparison of grabcut and f-g classification

set(num)	Weizmann 1obj(100)	Weizmann 2obj(100)	Grabcut dataset(50)
grabcut	0.85 ± 0.035	0.80 ± 0.046	0.89 ± 0.036
f-g classi.	**0.93 ± 0.010**	**0.90 ± 0.021**	**0.94 ± 0.016**

We also evaluate the method on the Berkeley segmentation dataset. Figure 3 gives some results. The adaptive initialization works well and generates good mean-shift patches. The minimum KL divergence D_K and the minimum Mahalanobis distance D_M make beneficial complements and greatly raise the chance of finding good segmentations.

117054(D_M) 361084(D_M) 370036(D_K) 361010(D_K)

Figure 3. Some segmentation results of the Berkeley image set.

The experiments reveal that the algorithm robustly propagates the background priors into the foreground mask region and *reliably treats multi-connectivity, multi-hole scenes*. As a typical example, almost all connected components and all holes in image 370036 are successfully identified. Such scenes are difficult for other schemes unless additional efforts are involved.

4. Conclusion

An adaptive figure-ground classification algorithm is proposed to do foreground extraction from bounding-box based background priors. The similarity measure is defined as the probability distance between adaptively generated mean-shift patched in a 5D feature space. Multiple hypotheses are employed to add the chance of success. This method achieves great success for multi-connectivity, multi-hole scenes.

A full paper with more details is given in reference [5].

References

[1] C. Rother, V. Kolmogorov, and A. Blake, "grabcut": interactive foreground extraction using iterated graph cuts. ACM Trans. Graph., 23(3):309–314, 2004.

[2] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. PAMI, 24(5):603–619, 2002.

[3] D. Comaniciu, An algorithm for data-driven bandwidth selection, PAMI, 25(2), 2003, pp. 281-288.

[4] J. Goldberger et al., An efficient image similarity measure based on approximations of KL-divergence between two gaussian mixtures, ICCV2003, vol. 1, pp. 487–493.

[5] Adaptive figure-ground classification, supplement materials.