Graph-based Joint Clustering of Fixations and Visual Entities

YUSUKE SUGANO, The University of Tokyo
YASUYUKI MATSUSHITA, Microsoft Research Asia
YOICHI SATO, The University of Tokyo

We present a method that extracts groups of fixations and image regions for the purpose of gaze analysis and image understanding. Since the attentional relationship between visual entities conveys rich information, automatically determining the relationship provides us a semantic representation of images. We show that, by jointly clustering human gaze and visual entities, it is possible to build meaningful and comprehensive metadata that offer an interpretation about how people see images. To achieve this, we developed a clustering method that uses a joint graph structure between fixation points and over-segmented image regions to ensure a cross-domain smoothness constraint. We show that the proposed clustering method achieves better performance in relating attention to visual entities in comparison with standard clustering techniques.

Categories and Subject Descriptors: H.1.2 [**Models and Principles**]: User/Machine Systems—*Human Information Processing*; I.4.9 [**Image Processing and Computer Vision**] Applications; I.5.3 [**Pattern Recognition**]: Clustering—*Algorithms*

General Terms: Human Factors

Additional Key Words and Phrases: Clustering, Image segmentation, ROI detection, Gaze analysis, Gaze visualization

ACM Reference Format:

Yusuke Sugano, Yasuyuki Matsushita, and Yoichi Sato. 2013. Graph-based Joint Clustering of Fixations and Visual Entities. *ACM Trans. Appl. Percept.* 10, 2, Article 1 (May 2013), 16 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

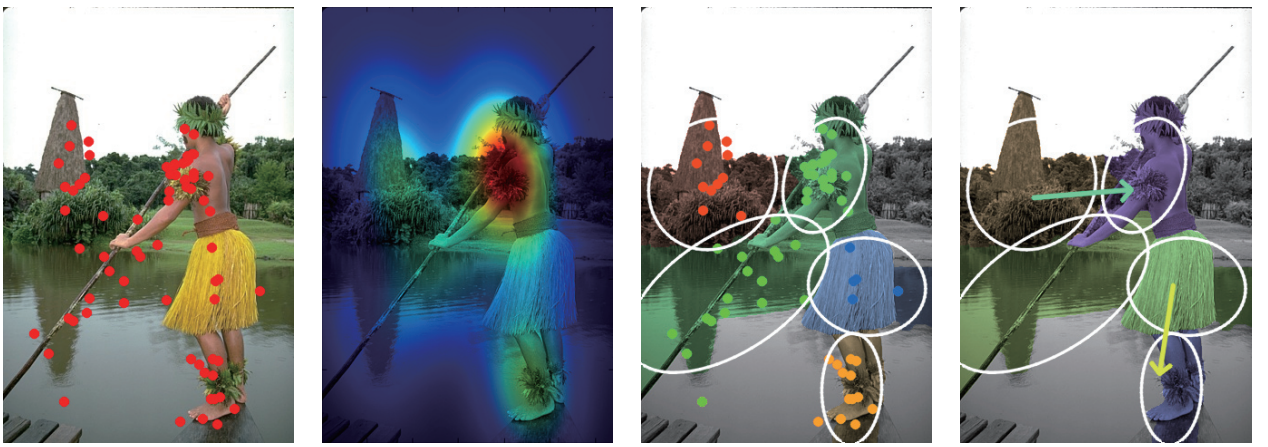
1. INTRODUCTION

Image understanding plays an important role in various applications, such as image search and retrieval, and hence, there has been a strong driving force for developing computer vision algorithms to derive semantic meanings from natural images. However, despite recent advances in the fields of object recognition and image understanding, it remains a difficult task for computers to interpret images as we see them.

The power of metadata in media understanding has gained a lot of attention with the recent explosive growth in the amount of online data. In real-world scenarios of media understanding, input media

This work is supported by CREST, JST. Author's address: Y. Sugano (corresponding author), Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505, Japan; email: sugano@iis.u-tokyo.ac.jp; Y. Matsushita, Microsoft Research Asia, No. 5 Dan Ling Street, Haidian District, Beijing, 100080, P. R. China; Y. Sato, Institute of Industrial Science, The University of Tokyo, 4-6-1 Komaba, Meguro-ku, Tokyo, 153-8505, Japan.

DOI: http://dx.doi.org/10.1145/0000000.0000000



(a) Original Fixation (b) Heatmap (c) Proposed Method (d) Transition Analysis

Fig. 1. Visualization of gaze data: (a) raw fixation data, (b) heat-map representation, (c) our joint clustering result, and (d) visualization of transition. The original image is adapted from [Arbeláez et al. 2011].

often has additional metadata such as user annotations and file tags. Since such metadata provides the context and semantic meaning of the media, it has been widely used for media understanding instead of directly tackling the difficult task of purely bottom-up image understanding. It is pointed out that, in various tasks such as recommending movies and tagging images, user-provided metadata plays the most important role, rather than multimedia content itself [Slaney 2011].

Subramanian *et al*. recently added a new perspective to the above scenario by discussing the possibility of using eye movement as metadata [Subramanian et al. 2011]. Human gaze information can be a unique cue for inferring our visual attention, and hence, it can provide knowledge about even *unconscious* visual context, which cannot always be given by literal metadata. They demonstrated the usage of the saccade data of a gaze for social and affective scene detection. The use of gaze information is becoming relevant as eye tracking techniques are getting mature [Hansen and Ji 2010], and low-cost consumer eye tracking devices are becoming more available, such as Mirametrix S2[1] and Tobii PC Eye[2]. As Subramanian *et al*. discussed, in the near future it will become possible to collect a large amount of gaze data on media contents, and therefore, establishing a method to utilize gaze data for image analysis and understanding is an important task.

For image analysis and understanding, it is important to analyze gaze data in relation to *visual entities*, *i.e*., image regions that correspond to the shapes of fixation target, and their regions of interest (ROIs). In contrast to a raw spatial distribution of the observer's attention often used by a heat-map representation (Fig. 1 (b)), ROI-based representation (Fig. 1 (c), (d)) conveys richer information about perceptual context. In the research fields such as experimental psychology and psychophysics, appropriateness of the ROI definition highly depends on experimental hypotheses [Holmqvist et al. 2011], and hence ROIs are manually defined in most cases. However, as discussed above, the demand and possibility of utilizing large-scale gaze data will further increase, and data mining approaches are expected to open a new vista on gaze analysis studies. It is accordingly required to explore an automatic and data-driven way of relating fixation clusters to semantically meaningful visual entities.

[1]http://mirametrix.com/products/eye-tracker/
[2]http://www.tobii.com/en/assistive-technology/global/products/hardware/pceye/

There are several technical challenges to achieving this goal. First, gaze locations recorded by eye trackers inevitably contain uncertainty due to both system errors and human eye jittering. Even commercial gaze trackers are reported to have errors of about 1.0 degree, and micro-saccades always occur around 2.0 degrees of the central visual field during fixations [Engbert 2006]. As a result, it becomes a non-trivial task to determine the exact image part that a person is looking at from a fixation location on the image. Second, automatic detection and grouping of visual entities in natural images (or generic object detection) has yet to be matured, partly due to ambiguous definition of visual entities. The ambiguity of grouping is also a fundamental issue in gaze analysis tasks as well, and this is another reason why manual fixation clustering and ROI definition are prefered [Holmqvist et al. 2011].

These two tasks are mutually related – clustering fixations and defining their corresponding visual entities. If there is a fixation cluster among multiple observers, it is likely to be associated with a visual entity – and vice versa. If relevant visual entities are defined over the image, the task of fixation clustering can benefit from this information. In [Subramanian et al. 2011], visual entities and their ROIs are first defined by object-specific detectors, and these ROIs are used for disambiguating the scale of the mean-shift clustering [Santella and DeCarlo 2004].

If the fixations are first clustered, the clusters help image segmentation/ROI definition tasks. Caldara and Miellet [Caldara and Miellet 2011] proposed a method to automatically define ROIs by analyzing the statistics of multiple fixation maps. Furthermore, Mishra *et al*. [Mishra et al. 2009] proposed a method that uses a single fixation point as a seed for graph-cut based image segmentation, and Subramanian *et al*. [Subramanian et al. 2010] extended their approach by using a cluster of multiple fixation points to achieve more stable segmentation results. Besides these gaze-focused studies, interactive image segmentation techniques have been widely studied in recent years [Boykov and Jolly 2001; Rother et al. 2004]. However, such methods often assume noise-free seed information that is manually provided by human annotators and cannot use noisy gaze information directly. Recently, Maji *et al*. [Maji et al. 2011] proposed a generic method that can incorporate such noisy and unreliable seed information by constraining the solution of a normalized cut to have correlation with the seed information. However, all these methods still cannot deal with the case of multiple objects where the gaze clusters are scattered, and the seeds have to be clustered into objects in advance.

We tackle the above chicken-and-egg problem by jointly clustering fixations and visual entities. Following the above discussion, we assume the structural similarity between fixation clusters and visual entities. In other words, we assume that the clustering result should form groups that have the similar structure in both fixation and image domains. Our method defines a graph structure that uses both fixation points and over-segmented image regions as graph nodes, and it uses the graph to exploit spatial correlation between fixation clusters and visual entities. The clustering task is accomplished through an iterative labeling process on the graph nodes through energy minimization. By considering the relationship between fixations and image regions, our joint approach produces a more comprehensive clustering than does independent clustering. In addition, unlike previous methods, our method can automatically determine the optimal number of clusters that are associated with groups of both fixations and visual entities. In this manner, our method (Fig. 1 (c)) tries to extract fixation clusters and corresponding visual entities in a fully unsupervised manner. By clustering fixations and relating them to visual entities, it becomes possible to analyze transitions among visual entities (Fig. 1 (d)) that are closely related to the semantic and perceptual context.

2. PROPOSED METHOD

In this section, the proposed method for joint clustering of fixations and visual entities is described. In our work, we assume that a fixation cluster forms a bivariate normal distribution in the image coordinates following the report [Nuthmann and Henderson 2010], which says fixation locations tend to

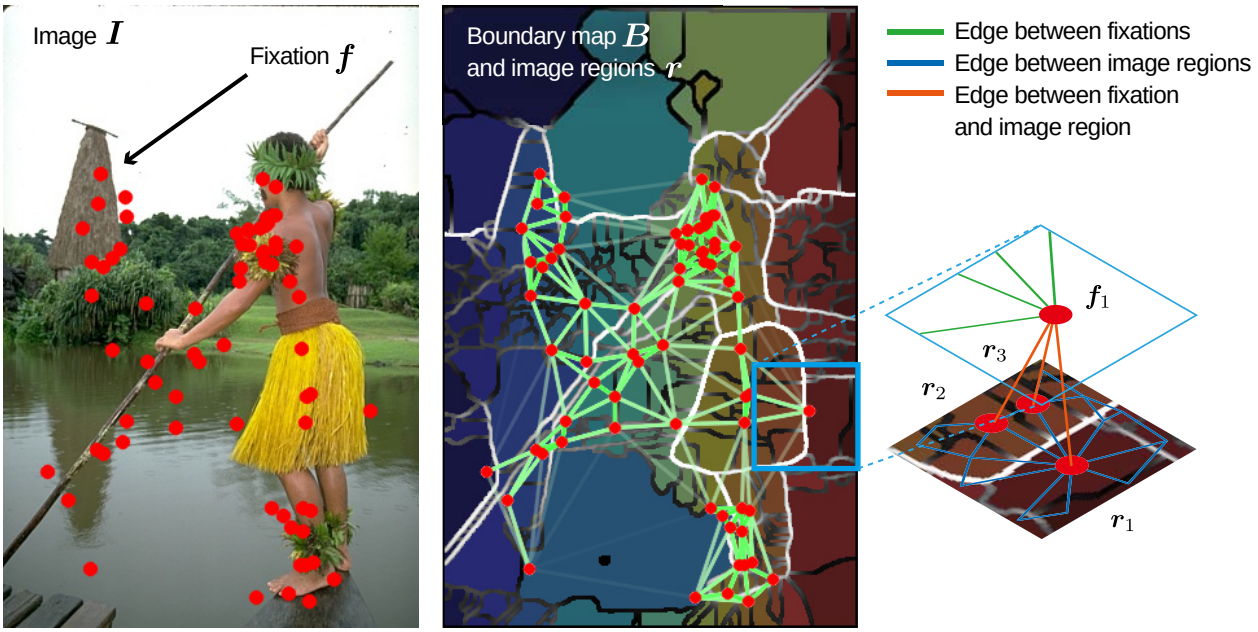


Fig. 2. Example of the graph structure computed by our method. The input to our method are a target image, $\boldsymbol{I}$, and fixations, $F = \{\boldsymbol{f}_n\}$. Over-segmented image regions $R = \{r_m\}$ associated with a boundary strength map, $\boldsymbol{B}$, are computed from $\boldsymbol{I}$, and our method uses both fixations F and over-segmented image regions R to form graph nodes. The original image is adapted from [Arbeláez et al. 2011].

make a normal distribution around the center of the object of interest. With this assumption, we cast the clustering problem as estimating the parameters of the normal distributions and finding the members (fixation points and image regions) of the clusters. To achieve the *joint* clustering, our method uses a joint graph structure defined over fixations and image regions and derives a solution via iterative energy minimization. At the heart of our joint clustering approach, connectivity of the fixations and image regions is defined over the graph in a unified manner, instead of ensuring the domain-specific connectivity (or smoothness) in fixation and image domains independently.

As depicted in Fig. 2, the input to our method is a target image $\boldsymbol{I}$ and N fixation locations $\{\boldsymbol{g}_n\}$ recorded from multiple people. From the target image $\boldsymbol{I}$, M over-segmented image regions $R = \{r_m\}$ associated with a *boundary strength map*, $\boldsymbol{B}$, is computed [Arbeláez et al. 2011]. Each of the over-segmented regions r_m corresponds to a set of pixels in the target image $\boldsymbol{I}$ that are in the same segment divided by the boundary map $\boldsymbol{B}$. A higher intensity in the boundary strength map $\boldsymbol{B}$ indicates a stronger object boundary. To deal with the noise in recorded gaze data, a set of fixations $F = \{f_n\}$ is represented by normal distributions around *recorded* fixation locations $\{\boldsymbol{g}_n\}$. That is to say, fixation f_n corresponds to a normal distribution with mean $\boldsymbol{g}_n$ and variance σ_f^2, which is set to a value corresponding to the expected system noise, *i.e.*, 1.0 degree in our setting. For notational consistency with $\{r_m\}$, we use $\{f_n\}$ to indicate a finite set of pixel locations in the image coordinates generated from the normal distribution. Our method uses both fixations F and over-segmented image regions R to form graph nodes. In what follows, we give more detailed definitions of the graph and joint clustering method for achieving our goal.

2.1 Graph Structure

Let us define two subsets of graph nodes, $\mathcal{V}_F$ and $\mathcal{V}_R$, where N nodes in $\mathcal{V}_F$ correspond to fixations in F, and M nodes in $\mathcal{V}_R$ correspond to image regions in R. In our method, a weighted graph, $G = (\mathcal{V}, \mathcal{E})$, is defined for the joint set of nodes $\mathcal{V} = \mathcal{V}_F \cup \mathcal{V}_R$. For graph edges $\mathcal{E}$, three types of edges are defined depending on the combination of nodes that are linked by the edge, *i.e.*, fixation-fixation edges, image region-image region edges, and fixation-image region edges. We describe how edge weights w are defined over nodes in the following.

Fixation-fixation edge.

To define edges between fixation nodes i and j ($i, j \in \mathcal{V}_F$), we use a Delaunay triangulation [De Berg et al. 2008] for the fixation locations. The triangulation is computed for the set of fixation locations $\{\boldsymbol{g}_n\}$ in the image coordinates, and an edge is defined between i- and j-th nodes if fixation nodes i and j are connected by the triangulation. The associated weight w_{ij} is defined so that it takes higher values as the two nodes are geometrically closer. A Gaussian function of the distance between node locations $\boldsymbol{g}_i$ and $\boldsymbol{g}_j$ is chosen as the weight function:

$$w_{ij} = \omega_a \exp\left(-\kappa ||\boldsymbol{g}_i - \boldsymbol{g}_j||^2\right), \tag{1}$$

where κ is a parameter of the Gaussian function.

Image region-image region edge. For edges between image region nodes $i, j \in \mathcal{V}_R$, an edge is defined between the nodes i and j if r_i and r_j are adjacent in the boundary map $\boldsymbol{B}$. As used in [Arbeláez et al. 2011], its weight w_{ij} is defined as a sigmoid function of the boundary strength. In our case, the weight is inversely correlated with the boundary strength as

$$w_{ij} = \omega_b \frac{1}{1 + \exp\left(\alpha(\bar{b} - \beta)\right)}, \tag{2}$$

where $\bar{b}$ indicates an average intensity of the boundary map $\boldsymbol{B}$ over the border between r_i and r_j. α and β are parameters of the sigmoid function.

Fixation-image region edge. If $i \in \mathcal{V}_F$ and $j \in \mathcal{V}_R$, a weight between the fixation node i and the image region node j is defined as an integral of the gaze distribution f_i over the image region r_i. In other words, the more completely the fixation distribution is contained in the image region, the more strongly the fixation node i is connected to the image region node j. As defined above, each f_i represents a normal distribution with mean $\boldsymbol{g}_i$ and variance σ_f^2, and hence, the integral can be computed as

$$w_{ij} = \omega_c \sum_{\boldsymbol{p} \in r_j} \frac{1}{\sqrt{2\pi}\sigma_f} \exp\left(-\frac{||\boldsymbol{p} - \boldsymbol{g}_i||^2}{2\sigma_f^2}\right), \tag{3}$$

where $\boldsymbol{p}$ denotes the pixel locations of the image region r_j in the image coordinates.

The defined edge weights w_{ij} are used for ensuring the connectivity of nodes. The higher weight w_{ij} has a greater cost in assigning different cluster labels to nodes i and j. Since the small weights do not contribute to the clustering task, edges with small weights are pruned by using a predefined threshold. In the above definitions of edge weights, ω_a, ω_b, and ω_c are used for controlling the relative strength of their contributions. If ω_a is set to a large value, for example, it requires a high cost to assign labels that are different from adjacent nodes, and therefore smoothness in the fixation domain is emphasized.

2.2 Objective Function

Now, we describe the objective function of the joint clustering. On the basis of the graph G defined in the previous subsection, the goal is to assign an optimal cluster label $l_i \in \mathcal{L}$ to each node $i \in \mathcal{V}$. Labels

$\mathcal{L}$ correspond to a finite set of a cluster hypothesis. As described earlier, each cluster hypothesis is modeled as a bivariate normal distribution, $\mathcal{N}(\boldsymbol{\mu}_{l_i}, \boldsymbol{\Sigma}_{l_i})$. Hence, each cluster label l_i is associated with its cluster parameters: mean $\boldsymbol{\mu}_{l_i}$ and covariance matrix $\boldsymbol{\Sigma}_{l_i}$.

The optimal labeling should maximize not only the suitability of the cluster hypotheses for fixation and image nodes but also smoothness over the graph G. The objective energy function $E(l)$ for the joint labeling l is defined as

$$E(l) = \sum_{i \in \mathcal{V}} D_i(l_i) + \sum_{(i,j) \in \mathcal{E}} V_{ij}(l_i, l_j) + \sum_{l^* \in \mathcal{L}} h_l \delta_{l^*}(l). \tag{4}$$

The first term in Eq. (4) indicates the data cost, which evaluates the appropriateness of assigning the label l_i to the node i. D_i is defined as

$$D_i(l_i) = \begin{cases} \sum_{\boldsymbol{p} \in f_i} p_{l_i}(\boldsymbol{p}) & \text{if } i \in \mathcal{V}_F \\ \sum_{\boldsymbol{p} \in r_i} p_{l_i}(\boldsymbol{p}) & \text{if } i \in \mathcal{V}_R \end{cases}, \tag{5}$$

where

$$p_{l_i}(\boldsymbol{p}) = -\log \left(\frac{1}{2\pi\sqrt{|\boldsymbol{\Sigma}_{l_i}|}} \exp \left(-\frac{1}{2} (\boldsymbol{p} - \boldsymbol{\mu}_{l_i})^T \boldsymbol{\Sigma}_{l_i}^{-1} (\boldsymbol{p} - \boldsymbol{\mu}_{l_i}) \right) \right) \tag{6}$$

is a log probability density function of the bivariate normal distribution corresponding to the cluster label l_i.

The second term of Eq. (4) indicates the smoothness cost, and V_{ij} is defined according to the edge weights as

$$V_{ij}(l_i, l_j) = \begin{cases} w_{ij} & \text{if } l_i \neq l_j \\ 0 & \text{otherwise} \end{cases}, \tag{7}$$

and the third term in Eq. (4) indicates the label cost that penalizes the total number of assigned unique labels:

$$\delta_{l^*}(l) = \begin{cases} 1 & {}^{\exists} i : l_i = l^* \\ 0 & \text{otherwise} \end{cases}, \tag{8}$$

and h_l is a constant value.

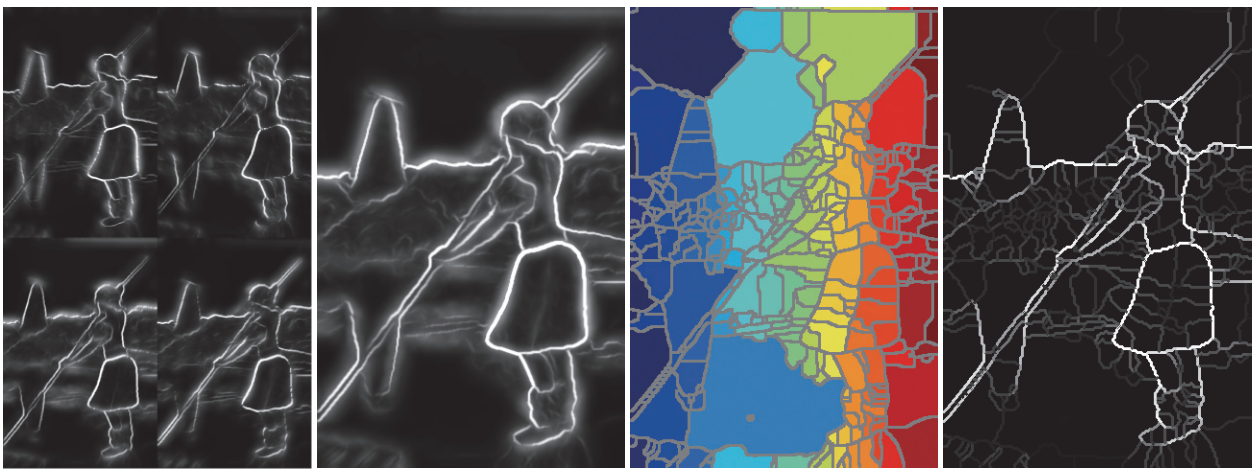
The objective function Eq. (4) can be efficiently minimized through an iterative energy optimization by using an α-expansion algorithm [Delong et al. 2012], and it yields optimal clustering of both fixation and image nodes.

3. IMPLEMENTATION DETAILS

In this section, we briefly give details on the above process: extraction of over-segmented image regions R and energy minimization of Eq. (4).

3.1 Image Segmentation

To extract a boundary strength map, $\boldsymbol{B}$, and image regions, R, from the input image $\boldsymbol{I}$, we use a gPb contour detection algorithm and oriented watershed transform [Arbeláez et al. 2011]. The contour map gPb is first computed from $\boldsymbol{I}$ by using a GPU-accelerated gPb algorithm [Catanzaro et al. 2009]. Their method is an approximation of the original gPb algorithm [Maire et al. 2008] that consists of local and global contour detectors. The local contour is detected on the basis of the oriented gradient strength. A multi-scale oriented signal, $mPb(\theta)$, at the angle θ is computed as a linear combination of the oriented gradient across different cues (brightness, color, and texture) and image scales. To enhance the global



(a) $gPb(\theta)$ (b) gPb (c) R (d) $\boldsymbol{B}$

Fig. 3. Boundary strength detection and image segmentation: (a) oriented contour maps $gPb(\theta)$, (b) contour map gPb, (c) over-segmented image regions R, and (d) boundary strength map $\boldsymbol{B}$. The original image is adapted from [Arbeláez et al. 2011].

structure of the local contour, the image region is then segmented into a few regions on the basis of the local contour map mPb, where $mPb(x, y) = \max_\theta mPb(x, y, \theta)$. In a similar manner to the normalized cuts-based image segmentation [Shi and Malik 2000], a pixel affinity matrix is defined as the maximum response of mPb along the line between two pixels, and segmented regions are computed via spectral clustering technique by using the affinity matrix. The global component $sPb(\theta)$ of the gPb algorithm is computed as the contours of the eigen-images, *i.e.*, solutions of the spectral clustering, with directional Gaussian derivative filters at angle θ. $gPb(\theta)$ (Fig. 3 (a)) is defined as a combination of the local and global components $mPb(\theta)$ and $sPb(\theta)$. The higher intensity in Fig. 3 (a) indicates a higher contour strength. As in the original implementation, $gPb(\theta)$ is computed in eight different orientations.

In the oriented watershed transform, the maximum contour strength $gPb = \max_\theta gPb(\theta)$ (Fig. 3 (b)) is first used for dividing $\boldsymbol{I}$ into over-segmented image regions R (Fig. 3 (c)) by using the standard watershed transform algorithm [Meyer 1994]. By treating a grayscale image as a topographical map, the catchment basins of local minima and their watershed lines are computed. Then, the boundary strength map $\boldsymbol{B}$ (Fig. 3 (d)) is defined in accordance with the orientation of the watershed lines. Watershed lines between image regions are approximated as straight lines, and their line orientations are quantized into eight orientations as $gPb(\theta)$. Finally, $\boldsymbol{B}$ is defined by assigning the value of $gPb(\theta)$ of the corresponding orientation θ to watershed lines.

3.2 Energy minimization

The objective function in Eq. (4) is defined on the discrete labeling l; however, in our case, each cluster label is associated with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, which are in a continuous space. Hence, as discussed by Delong *et al.* [Delong et al. 2012], it is necessary to iteratively update both labeling l and cluster parameters as follows.

Algorithm 1 summarizes our iterative clustering method. The initial set of cluster hypotheses $\mathcal{L}_0$ is built so that each fixation f_n constructs an independent cluster. For each node $i \in \mathcal{V}_F$, a unique label, l_i, is assigned with parameters $\boldsymbol{\mu}_{l_i} = \boldsymbol{g}_i$ and $\boldsymbol{\Sigma}_{l_i} = \begin{bmatrix} \sigma_f^2 & 0 \\ 0 & \sigma_f^2 \end{bmatrix}$. In addition, a special background label ϕ

ALGORITHM 1: Iterative Clustering

1: Initialize cluster hypotheses $\mathcal{L}_0$
2: **repeat**
3: Compute optimal labeling l w.r.t. $\mathcal{L}_t$ by minimizing Eq. (4) by using the α-expansion algorithm
4: Re-estimate cluster parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ to obtain $\mathcal{L}_{t+1}$
5: **until** Convergence

with a constant data cost,

$$D_i(\phi) = \begin{cases} \infty & \text{if } i \in \mathcal{V}_F \\ \epsilon & \text{if } i \in \mathcal{V}_R \end{cases} , \tag{9}$$

which allows only image regions to take a small constant value ϵ, is assigned to nodes $i \in \mathcal{V}_R$.

Given a set of discrete labels, the optimal labeling l can be computed by minimizing Eq. (4) by using the modified α-expansion algorithm [Delong et al. 2012]. Then, the parameters of each cluster hypothesis are updated. Since cluster parameters only affect the data cost D_i, Eq. (4) can be further minimized by setting $\boldsymbol{\mu}_{l_i}$ and $\boldsymbol{\Sigma}_{l_i}$ as the mean and the covariance of the associated points. In this manner, Eq. (4) is minimized via iteration until convergence. If the same label is assigned to both fixations and image regions, they belong to the same cluster.

4. EXPERIMENTS

We conducted experiments to evaluate the performance of our method by using the human-annotated ground truth data of clustering. Ideally, we would like to evaluate the performance directly; however, since there has been no method that can perform joint clustering equivalent to that of our method, the accuracy of fixation clustering and image segmentation were separately compared with standard clustering methods.

To construct a test data set and its ground-truth annotations, we used a BSDS500 data set [Arbeláez et al. 2011]. Ten novice human subjects were first used for recording fixation locations on images with a Tobii TX300 Eye Tracker[3]. Images were displayed on the built-in 23-inch display of the TX300 tracker, while test subjects fixed their head positions 65 [cm] away from the display. Each image was shown for 4 seconds in a random order, and a white cross mark on a black background was displayed at the display center for 2 seconds between images to capture human attention. Gaze data were recorded at 60 [Hz], and they were divided if the gaze acceleration exceeded a threshold of 6 [deg/sec], and the median location was used for fixation $\boldsymbol{g}$ for each of the divided clusters of gaze data[4].

The scaling parameters of the edge weight functions are empirically set at $\omega_a = 90$, $\omega_b = 30$, $\omega_c = 200$, while the label cost is set at $h_l = 300$. Although the computational cost depends on the number of graph nodes N and M, it took about 2 seconds per one image for energy optimization using a 3.33-GHz Core i7 CPU in our current implementation.

Fig. 4 shows some examples from the data set and our clustering results. Each overlaid circle indicates fixation locations, and the colors indicate assigned cluster labels of both fixations and image regions. If some labels were assigned only to fixations and no corresponding visual entity was found, these fixation points were considered as outliers and discarded before the visualization. As we can observe, fixation clusters were robustly found across images, and their corresponding regions of visual entities were also well segmented out from background regions.

[3]http://www.tobii.com/
[4]The dataset is available at http://www.hci.iis.u-tokyo.ac.jp/datasets/

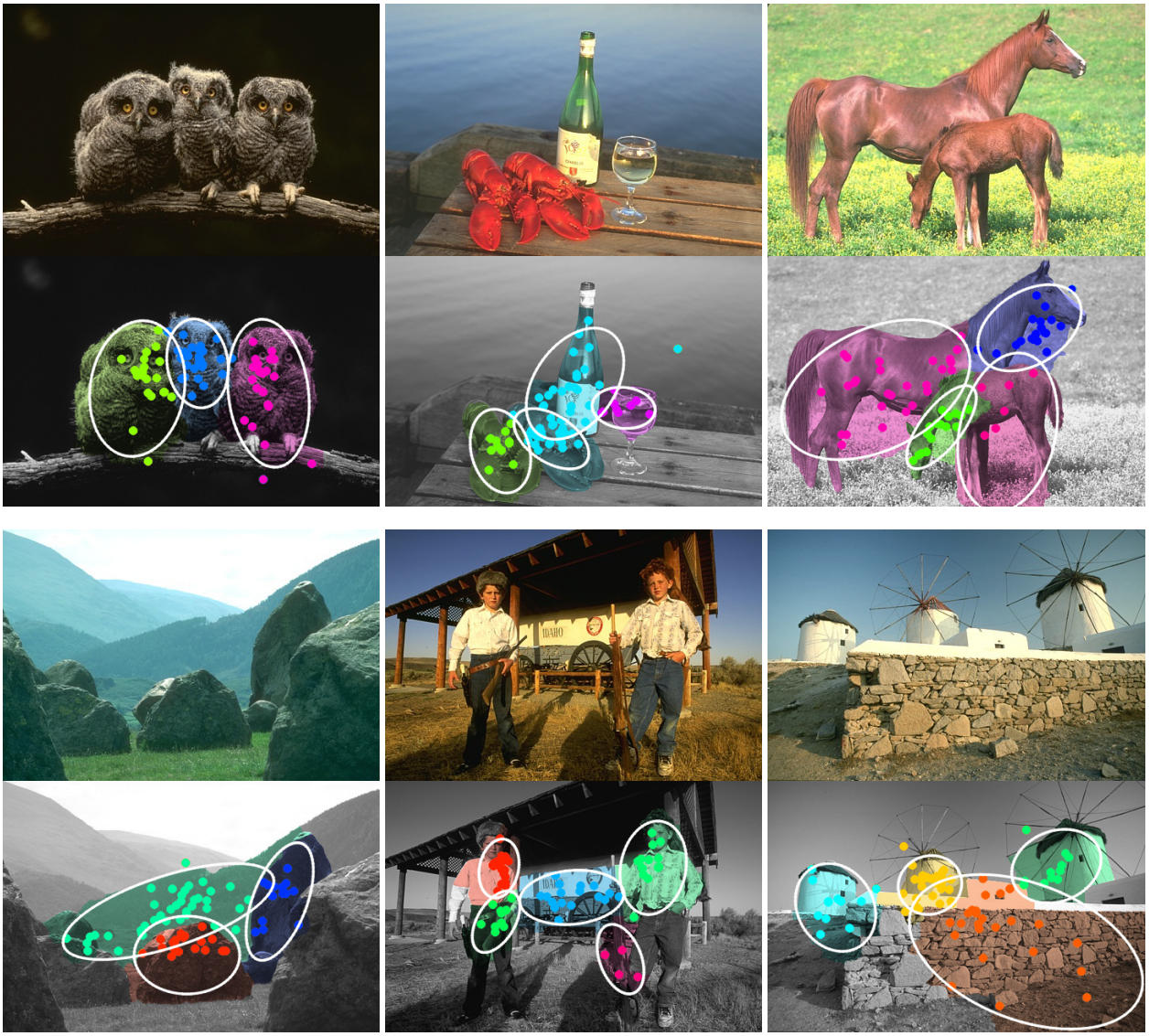


Fig. 4. Examples of clustering results. Overlaid circles indicate fixation locations, and the colors represent assigned cluster labels of both fixations and image regions. The original images are adapted from [Arbeláez et al. 2011].

At the same time, five different human subjects were asked to give ground-truth annotations on 100 images chosen from the above test data set. The task for the subjects was to assign unique labels to sets of fixations and image regions. For each of the 100 images, a fixation-overlaid color image and a gray-scale contour image were displayed side by side to the subjects as shown in Fig. 5. Since the BSDS500 data set also contains human annotations of object boundaries, aggregated annotation maps were used as contour maps instead of the automatically extracted ones. Subjects were asked to use the mouse cursor to lasso the fixations and to paint the contour image into one cluster at a time. The only instruction given to the subjects was to make *semantically meaningful* clusters, and the total number

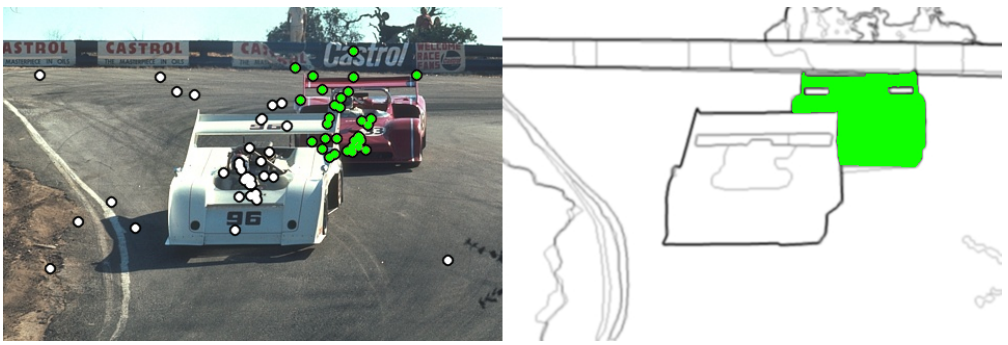


Fig. 5. Image display for ground-truth annotation. A fixation-overlaid color image and a gray-scale contour image were displayed side by side, and the subjects were asked to use the mouse cursor to lasso the fixations and to paint the contour image into one cluster at a time. The original image is adapted from [Arbeláez et al. 2011].

of clusters was left to their decision. It was also allowed to leave some fixation points unlabeled to indicate they are outlier points.

For the purpose of comparison, Gaussian mixture model (GMM) was fitted to the fixation data as a baseline result for fixation clustering [Pedregosa et al. 2011]. Additionally, k-means clustering was selected from commonly used fixation clustering methods as another baseline method. Image regions were produced using the ultra-metric contour maps (UCMs) [Arbeláez 2006], *i.e.*, hierarchical representation of object boundaries, provided together with the BSDS500 data set. Since the performance of standard clustering methods highly depend on parameters, the number of clusters k for these baseline clustering methods was set to the mean number of clusters given by the five test subjects. UCMs are thresholded at the highest value that produces k image regions. The number k was given to only baseline methods, but not to the proposed method. Parameters of the proposed method was empirically defined and fixed through experiments.

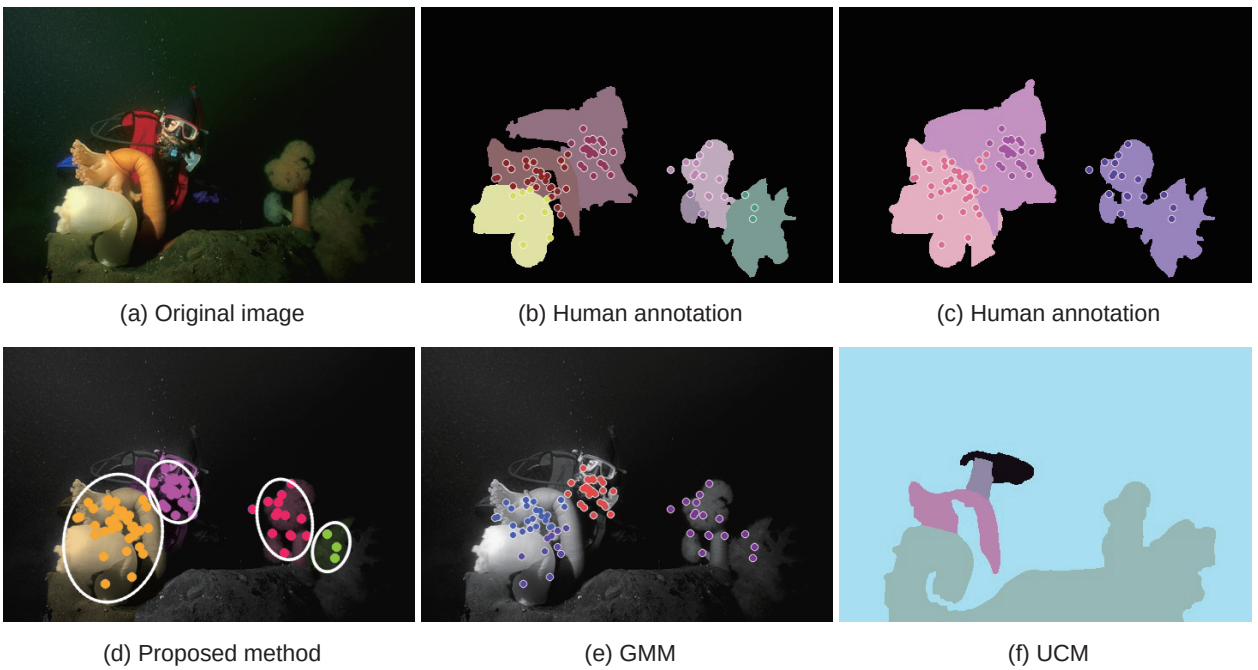
Fig. 6 shows typical examples of clustering results compared against human annotation. Although cluster labels were not always consistent between human annotators (Fig. 6 (b), (c)), it can be seen that the proposed method (Fig. 6 (d)) achieved a clustering result similar to humans. Given the number of clusters, GMM fixation clustering (Fig. 6 (e)) could often achieve a clustering result similar to our method; however, it sometimes created clusters that were not consistent with visual entities. Fully unsupervised image segmentation based on UCM (Fig. 6 (f)) is fundamentally a difficult task, and visual entities were not clearly extracted.

The accuracy of the clustering was assessed on the basis of two metrics: Hubert-Arabie adjusted Rand index [Hubert and Arabie 1985] and V-measure [Rosenberg and Hirschberg 2007]. The adjusted Rand index (ARI) is a chance-adjusted version of the original Rand measure that is defined to take 0 for a completely random clustering result. Given a set of ground-truth cluster labels L_g and estimated cluster labels L_e of S elements, the original Rand index [Rand 1971] is computed as

$$RI = \frac{n_a + n_b}{\binom{S}{2}}, \tag{10}$$

where n_a is the number of element pairs that have the same label in both L_g and L_e, and n_b is the number of element pairs that have different labels in both L_g and L_e. The ARI is defined on the basis of the above definition by following the general form of chance correction:

$$ARI = \frac{RI - Expected(RI)}{Max(RI) - Expected(RI)}, \tag{11}$$



(a) Original image (b) Human annotation (c) Human annotation

(d) Proposed method (e) GMM (f) UCM

Fig. 6. Examples of clustering results. From left to right, top to bottom: (a) original input image, (b), (c) human annotations of two different human subjects, (d) joint clustering results of fixations and image regions with the proposed method, (e) fixation clustering with GMM, and (f) image segmentation by UCM. The original image is adapted from [Arbeláez et al. 2011].

where $Max(RI)$ and $Expected(RI)$ are the maximum and the expectation of the Rand index given L_g and L_e, respectively.

While the ARI is commonly used for evaluating clustering accuracy, one of the biggest drawbacks is that it lacks intuitiveness, and ARI scores cannot be compared qualitatively. For that reason, we additionally used the V-measure, which is defined as a harmonic mean of homogeneity h and completeness c for the evaluation:

$$V_\beta = (1+\beta)\frac{hc}{\beta h + c}, \tag{12}$$

where the weight of homogeneity β is set to 1 in our case. Homogeneity h and completeness c are defined in an information-theoretic manner as

$$h = 1 - \frac{H(L_g|L_e)}{H(L_g)} \quad \text{and} \quad c = 1 - \frac{H(L_e|L_g)}{H(L_e)}. \tag{13}$$

$H(L_g)$ is the entropy of the ground-truth cluster labels L_g, and $H(L_g|L_e)$ is the conditional entropy of L_g given L_e. Homogeneity h becomes higher if the estimated labels contain only elements that are assigned to the same cluster in the ground-truth labeling. Completeness c is symmetrically defined, and it takes a higher value if the estimated labels contain all elements belonging to the same cluster in the ground-truth labeling.

Table I and Fig. 7 summarize the mean scores of 5 subjects $\times 100$ images for the ARI and V-measure. In Fig. 7, the mean scores for homogeneity and completeness are additionally shown. Image segmentation is still a difficult task even with the number of clusters given. The proposed method achieved significantly better scores in the ARI (paired t-test: $t(499) = 14.34$, $p < 0.01$, effect size $r = 0.54$) and

Table I. Comparison of ARI and V-measure scores (mean ± standard deviation) between baseline and proposed methods

	Fixations			Image Regions	
	GMM	k-means	**Proposed**	UCM	**Proposed**
ARI	0.42 ± 0.28	0.45 ± 0.30	$\mathbf{0.48 \pm 0.30}$	0.63 ± 0.14	$\mathbf{0.74 \pm 0.13}$
V-measure	0.50 ± 0.26	0.51 ± 0.29	$\mathbf{0.54 \pm 0.28}$	0.26 ± 0.19	$\mathbf{0.39 \pm 0.13}$

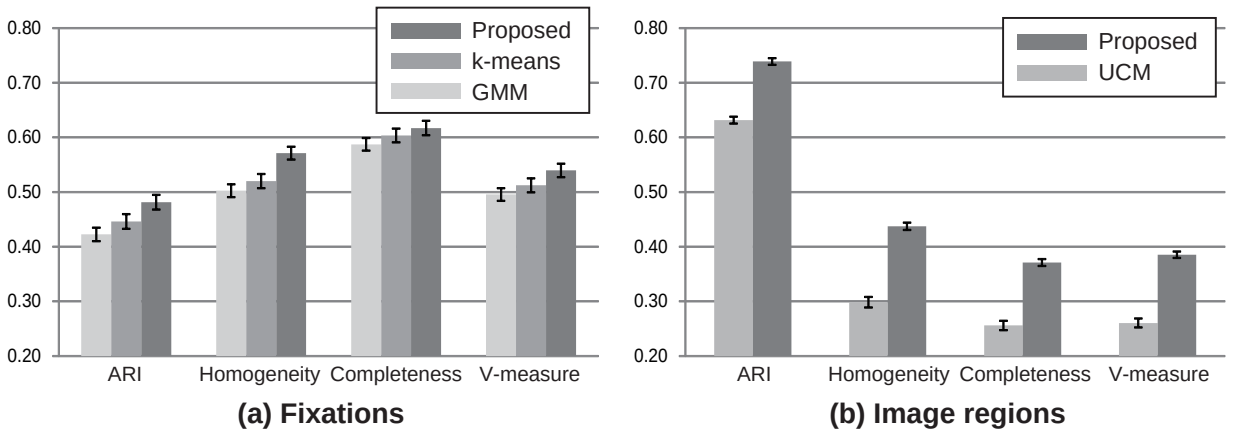


Fig. 7. Comparison of clustering accuracies. Each bar indicates mean score for ARI and V-measure (homogeneity and completeness are additionally shown). The light bars correspond to baseline methods, and the dark bars correspond to the proposed method. The error bars indicate standard errors.

V-measure (paired t-test: $t(499) = 12.50$, $p < 0.01$, $r = 0.49$). In terms of fixation clustering, even simple clustering methods like GMM fitting and k-means can have a good chance of achieving an accurate result by specifying the number of clusters. However, it can be seen that the proposed method achieved higher scores than did the baseline methods despite the fact that the true number of clusters was not provided in our method. Although their effect sizes are not significantly large, our proposed method achieved better score in both ARI (paired t-test: GMM, $t(499) = 5.24$, $p < 0.01$, $r = 0.23$ and k-means, $t(499) = 3.12$, $p < 0.01$, $r = 0.14$) and V-measure (paired t-test: GMM, $t(499) = 4.60$, $p < 0.01$, $r = 0.20$ and k-means, $t(499) = 2.86$, $p < 0.01$, $r = 0.13$).

4.1 Transition Visualization

As discussed earlier, joint clustering of fixations and visual entities enables us to further analyze transition statistics between visual entities. In this section, we show some examples of transition visualization that can be subsequently achieved with the proposed method. In the following figures, gaze data and images are taken from [Judd et al. 2009].

While there have been several methods for analyzing transitions [Holmqvist et al. 2011], we employed Subramanian *et al.*'s formulation [Subramanian et al. 2010]. They proposed using saccade likelihood $P(l_j|l_i)$ from l_i to l_j to analyze transition statistics:

$$P(l_j|l_i) = \frac{n_s(l_i, l_j)}{n_f(l_i)}, \tag{14}$$

where $n_s(l_i, l_j)$ is the number of saccades from l_i to l_j, and $n_f(l_i)$ is the number of fixations in l_i. In the following examples, arrowed lines connecting clusters indicate saccade likelihoods. The value $P(l_j|l_i)$

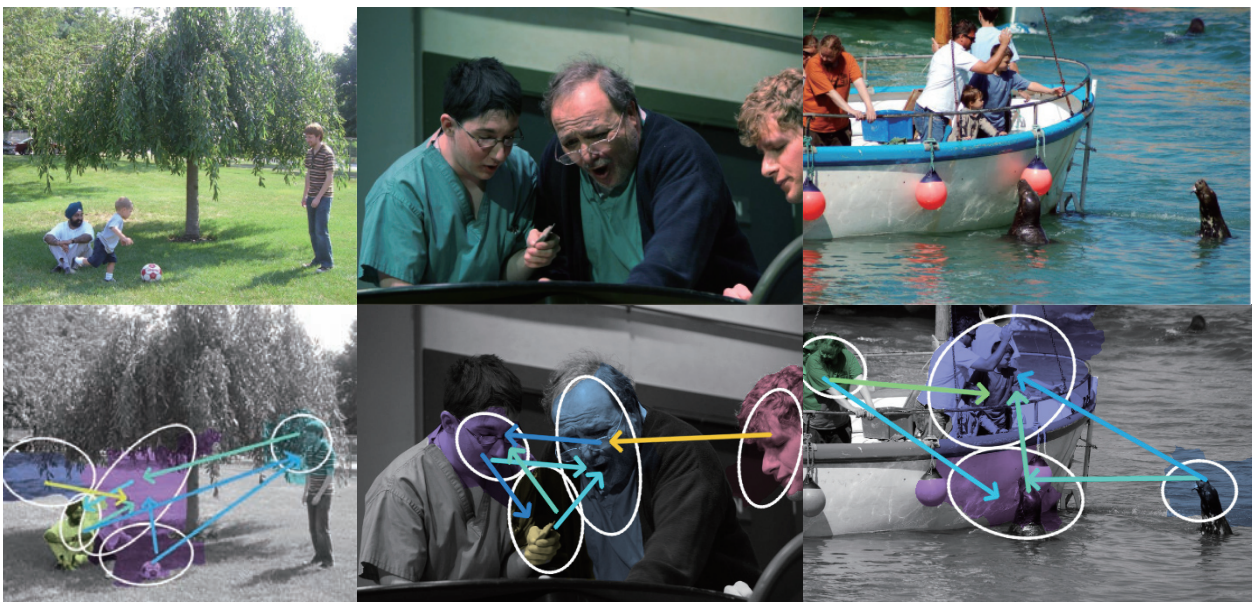


Fig. 8. Examples of transitions between interacting objects. The original images are adapted from [Judd et al. 2009].

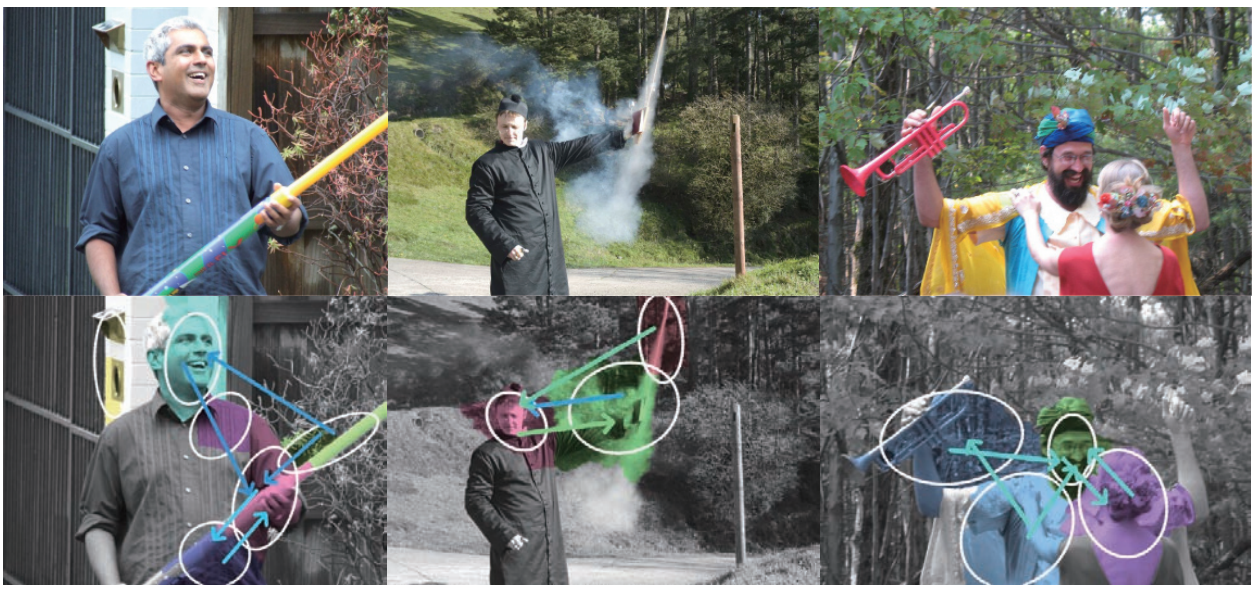


Fig. 9. Examples of transitions between face and hand-held object. The original images are adapted from [Judd et al. 2009].

is encoded by using a Jet color-map from blue ($P = 0.0$) to red ($P = 1.0$); however, ϕ lower than a threshold ($\phi < 0.25$) are discarded in the following figures for clearer presentation.

As reported in [Subramanian et al. 2010], frequent transitions can often be observed between interacting or semantically related objects, as shown in Fig. 8. Not only between interacting persons, semantic relationships with animals or non-living objects can cause gaze transitions. In the images

Fig. 10. Examples of transitions on large sign-boards. The original images are adapted from [Judd et al. 2009].

Fig. 11. Examples of transitions between head and body. The original images are adapted from [Judd et al. 2009].

shown in Fig. 9, frequent transitions happened between the face and the object the person held. These examples show another interesting case of transitions that indicates a semantic relationship of the image contents.

In particular, when the object of interest is large in the image, there tend to be several fixation clusters within the object. Another typical category of frequent transitions is the ones between such

intra-object clusters. Fig. 10 clearly demonstrates intra-object transitions on sign boards, and transitions between head and body as shown in Fig. 11 are also a commonly observed case.

As demonstrated in these examples, the proposed method can provide a way to analyze transition statistics in a fully bottom-up manner without depending on predefined ROIs. This will enable easier handling of large scale datasets, leading to further understanding on how eye movements can contribute to image understanding tasks.

5. SUMMARY

In this work, we proposed a novel graph-based joint clustering method for fixations and visual entities. On the basis of the observation that there is correspondence between fixations and image clusters, a weighted graph was built to ensure global smoothness jointly on fixations and image regions. Joint clustering is done through an iterative energy optimization procedure over a graph, and this leads to more robust and accurate results than does separate clustering.

The proposed method was shown to be able to achieve a higher clustering accuracy than do standard separate clustering methods. It can robustly discover the structure of attention on a large set of images and fixations, and it will open a new way to perform web-scale multimedia analysis, as claimed in [Subramanian et al. 2011]. Our future work includes its applications on scene analysis and image organization tasks.

REFERENCES

P. Arbeláez. 2006. Boundary extraction in natural images using ultrametric contour maps. In *Proc. 5th IEEE Workshop on Perceptual Organization in Computer Vision (POCV'06)*. 182–182.

P. Arbeláez, M. Maire, C. Fowlkes, and J. Malik. 2011. Contour Detection and Hierarchical Image Segmentation. http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/resources.html, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 5 (2011), 898–916.

Y. Boykov and M.-P. Jolly. 2001. Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In *Proc. 8th IEEE International Conference on Computer Vision (ICCV 2001)*. 105–112.

Roberto Caldara and Sebastien Miellet. 2011. iMap: a novel method for statistical fixation mapping of eye movement data. *Behavior Research Methods* 43 (2011), 864–878. Issue 3.

Bryan Catanzaro, Bor-Yiing Su, Narayanan Sundaram, Yunsup Lee, Mark Murphy, and Kurt Keutzer. 2009. Efficient, high-quality image contour detection. In *Proc. 12th IEEE International Conference on Computer Vision (ICCV 2009)*. 2381–2388.

M. De Berg, O. Cheong, M. Van Kreveld, and M. Overmars. 2008. *Computational geometry: algorithms and applications*. Springer.

Andrew Delong, Anton Osokin, Hossam Isack, and Yuri Boykov. 2012. Fast Approximate Energy Minimization with Label Costs. *International Journal of Computer Vision* 96 (2012), 1–27. Issue 1.

Ralf Engbert. 2006. Microsaccades: a microcosm for research on oculomotor control, attention, and visual perception. In *Visual Perception Fundamentals of Vision: Low and Mid-Level Processes in Perception*. Progress in Brain Research, Vol. 154, Part A. Elsevier, 177 – 192.

D. W. Hansen and Qiang Ji. 2010. In the Eye of the Beholder: A Survey of Models for Eyes and Gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 3 (2010), 478 –500.

K. Holmqvist, M. Nyström, R. Andersson, R. Dewhurst, H. Jarodzka, and J. Van de Weijer. 2011. *Eye tracking: A comprehensive guide to methods and measures*. Oxford University Press.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification* 2 (1985), 193–218. Issue 1.

T. Judd, K. Ehinger, F. Durand, and A. Torralba. 2009. Learning to Predict Where Humans Look. http://people.csail.mit.edu/tjudd/WherePeopleLook/index.html. In *Proc. Twelfth IEEE International Conference on Computer Vision (ICCV 2009)*.

M. Maire, P. Arbeláez, C. Fowlkes, and J. Malik. 2008. Using contours to detect and localize junctions in natural images. In *Proc. IEEE Conference on Computer Vision & Pattern Recognition (CVPR 2008)*. 1–8.

S. Maji, N. K. Vishnoi, and J. Malik. 2011. Biased normalized cuts. In *Proc. IEEE Conference on Computer Vision & Pattern Recognition (CVPR 2011)*. 2057–2064.

Fernand Meyer. 1994. Topographic distance and watershed lines. *Signal Processing* 38, 1 (1994), 113 – 125.

Ajay Mishra, Yiannis Aloimonos, and Cheong Loong Fah. 2009. Active segmentation with fixation. In *Proc. 12th IEEE International Conference on Computer Vision (ICCV 2009)*. 468–475.

Antje Nuthmann and John M. Henderson. 2010. Object-based attentional selection in scene viewing. *Journal of Vision* 10, 8 (2010).

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

William M. Rand. 1971. Objective Criteria for the Evaluation of Clustering Methods. *J. Amer. Statist. Assoc.* 66, 336 (1971), 846–850.

A. Rosenberg and J. Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proc. 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. 410–420.

Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. 2004. "GrabCut": interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics* 23, 3 (2004), 309–314.

Anthony Santella and Doug DeCarlo. 2004. Robust clustering of eye movement recordings for quantification of visual interest. In *Proc. 2004 Symposium on Eye tracking research & applications (ETRA '04)*. 27–34.

J. Shi and J. Malik. 2000. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 8 (2000), 888–905.

M. Slaney. 2011. Web-Scale Multimedia Analysis: Does Content Matter? *IEEE Multimedia* 18, 2 (2011), 12–15.

Ramanathan Subramanian, Harish Katti, Nicu Sebe, Mohan Kankanhalli, and Tat-Seng Chua. 2010. An Eye Fixation Database for Saliency Detection in Images. In *Proc. 11th European Conference on Computer Vision (ECCV 2010) (Lecture Notes in Computer Science)*, Vol. 6314. 30–43.

Ramanathan Subramanian, Victoria Yanulevskaya, and Nicu Sebe. 2011. Can computers learn from humans to see better?: inferring scene semantics from viewers' eye movements. In *Proc. 19th ACM International Conference on Multimedia (MM '11)*. ACM, 33–42.