This is the author's version of an article that has been published in this journal. Changes were made to this version by the publisher prior to publication.
The final version of record is available at http://dx.doi.org/10.1109/TIP.2018.2883580


# A New Probabilistic Representation of Color Image Pixels and Its Applications

Zhouchi Lin, Hongdong Qin, and S. C. Chan, *Member, IEEE*[1]

***Abstract*— This paper proposes a novel probabilistic representation of color image pixels (PRCI) and investigates its applications to similarity construction in motion estimation and image segmentation problems. The PRCI explores the mixture representation of the input image(s) as prior information and describes a given color pixel in terms of its membership in the mixture. Such representation greatly simplifies the estimation of the probability density function from limited observations and allows us to derive a new probabilistic pixel-wise similarity measure based on the continuous domain Bhattacharyya coefficient. This yields a convenient expression of the similarity measure in terms of the pixel memberships. Furthermore, this pixel-wise similarity is extended to measure the similarity between two image regions. The usefulness of the proposed pixel/region-wise similarities is demonstrated by incorporating them respectively in a dense image descriptor-based multi-layered motion estimation problem and an unsupervised image segmentation problem. Experimental results show that i) the integration of the proposed pixel-wise similarity in dense image-descriptor construction yields improved peak signal to noise ratio performance and higher tracking accuracy in the multi-layered motion estimation problem, and ii) the proposed similarity measures give the best performance in terms of all quantitative measurements in the unsupervised superpixel-based image segmentation of the MSRC and BSD300 datasets.**

***Index Terms*—Probabilistic color representation, Pixel-wise similarity, Region-wise similarity, Image matching, registration, and segmentation, Image descriptors.**

## I. Introduction

COLOR of natural images is usually represented by various color spaces such as RGB, Lab, YCbCr, etc. in form of a three dimensional vector. In many image processing tasks, it is often required to measure the similarity of two image locations or regions. For instance, in image registration, a reference image may be warped to another target image. Therefore, it is required to measure how close the warped pixels are to their corresponding pixels in the target. The color component, such as the RGB color space, is valuable to quantify this similarity. In particular, the Euclidean norm of the difference of the two color components is a commonly used measure in measuring their similarity/difference. Color components also play a vital role in image segmentation. Due to the large variability of image content and corruption due to noise, it is also desirable to be able to construct more reliable similarity measure for differentiating two image regions.

Zhouchi Lin, Hongdong Qin, and S. C. Chan are with the Department of Electrical and Electronic Engineering, the University of Hong Kong (e-mail: {zclin; hdqin; scchan@eee.hku.hk}).

This paper has supplementary downloadable material available at https://www.eee.hku.hk/~dsp/download/Supplementary_PRCI.pdf, provided by the author. The material includes additional experimental results as well as related proofs of certain equations and statements. Contact [zclin; hdqin; scchan@eee.hku.hk] for further questions about this work.

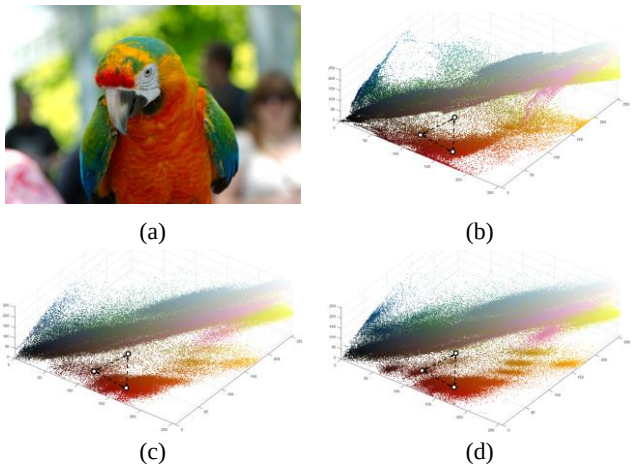

Figure 1. (a). An example image; (b) Color pixels of (a) in RGB color space; (c) Random samples generated by Gaussian Mixture components estimated from (b); (d) Random samples generated by Laplace Mixture components estimated from (b). The three small circles denote three colors with the same Euclidean distance from each other, but the two at the bottom are more closely related in terms of image content. This suggests that the Euclidean distance is not a good similarity measure. Compared to (c), (d) is more concentrated at the center of each cluster but with a larger spread. This suggests that the Laplacian mixture may be a more realistic model of natural images.

While the Euclidean norm of color components may seem to be a natural similarity measure, previous research has shown that it may not be the best possible measure. For instance, in computational color naming [32-35], one aims to build a relationship between color pixel values to color name labels/vocabularies. It was found in the Berlin and Kay's experiment [36] that human can perform much better in picking the 'best example' for each of the color terms than in establishing boundaries between the prototypical colors. This suggests that human may prefer working on a group of representative colors rather than measuring precisely the continuous variations in the color space. This motivates us to develop a probabilistic approach to describe the color distribution of an image or a group of images and derive new pixel-wise and region-wise similarities from such a representation. Previous studies have shown that an arbitrary color value may be represented in form of a graded (fuzzy) membership of prototypical colors [33-35] to facilitate contextual reasoning. Moreover, the 11-dimension membership of color name feature proposed by Joost van de Weijer [33] offers more linguistic information than the original 3-dimensional color space representation.

The potential advantages of having a probabilistic representation of the color pixels in an image or a set of images can be visualized from Fig. 1. Fig. 1(a) and Fig 1(b) show respectively the original image of a parrot and its corresponding color distribution. Since the distribution is not uniformly distributed, the deterministic Euclidean distance

The final version of record is available at http://dx.doi.org/10.1109/TIP.2018.2883580


between two color pixels may not be a good measure of their similarity. For instance, the three black points in Fig. 1(b) to (d) are RGB colors with the same Euclidean distance to each other, but one of the points does not correspond to any popular color in the image and hence its similarity with the other two should be much smaller. This suggests that the Euclidean distance is not a good similarity measure. Moreover, as the pixels seem to cluster around special regions, it is more efficient to represent them in form of a mixture of component distributions. Since the component distributions will depend on the image or image sequences, they should be estimated from the image data so that they can serve as useful prior information to define the membership of a given color pixel in the component distribution, and hence their distance and similar measures. The estimation of this global finite mixture using Gaussian component density can be done by means of expectation- maximization (EM) algorithm [7]. However, the results usually depend on a good initialization of the mixture density. Moreover, an efficient algorithm which can determine the number of clusters is highly desirable to avoid excessive complexity. Fig. 1(c) shows an estimated mixture with Gaussian component density. It can be seen that it resembles quite closely the original distribution except that the Gaussian distribution usually decay quite rapidly. On the other hand, unlike the original image, the distribution of color pixels generated by the Laplace mixture proposed in this paper in Fig.1(d) is more concentrated at the cluster center and having a larger spread. This suggests that the Laplacian mixture may be a more realistic model of natural images.

In addition, using this global color components as prior information, one can express a given pixel equivalently as the membership of these component densities. Pixels coming from two different component distributions may therefore possess low similarity as compared with those inside the same component. Moreover, the Euclidean distance between the two pixels coming from different components may not be very large. This suggests that the membership representation has higher discrimination power than the Euclidean distance due to the prior information of the color distribution.

Furthermore, if the mixture components learned from the image(s) is used as the prior information, then the estimation of the local probability density function (pdf) of an image region can be significantly simplified. In fact, instead of estimating directly the location, spread and mixing weights of the mixture components of the associated local pdf, it is sufficed to estimate the mixing weights, which will require considerably less number of samples. As the local pdf can be conveniently estimated from the memberships of the observed pixels, it suggests that more fundamental probabilistic region-wise similarity may be developed from the estimated pdf and hence the membership representation, say using the Bhattacharyya coefficient (BC) [9]. Consequently, such a representation may benefit a wide range of image processing applications.

In this paper, we further develop these observations and propose a new probabilistic representation of color image pixels (PRCI) which consists of the membership of a given image pixel in a finite mixture of multivariate Gaussian as well as Laplace component densities. Based on this component representation of color pixels, we then propose a pixel-wise similarity measure of two given pixels by approximating the continuous domain Bhattacharyya coefficient (BC)[2]. This yields a convenient expression in terms of their memberships in the finite component mixture. Moreover, we shall show in the supplementary material that fewer samples are needed to estimate the local pdf from samples than direct estimation using K-means, given the prior mixture components. Thus, the PRCI can better capture the local pdf by exploring the prior information of the mixture components. Furthermore, this pixel-wise similarity is extended to measure the similarity between two image regions.

The usefulness of the proposed pixel-wise and region-wise similarity is demonstrated by incorporating them respectively in a dense image descriptor-based multi-layered motion estimation problem and an unsupervised image segmentation problem. Experimental results show that i) the proposed similarity yields improved peak signal to noise ratio (PSNR) performance and higher tracking accuracy in terms of Dice Coefficient [21] over the state-of-the-art dense Scale- and Rotation-Invariant descriptor (SID) [12] and ii) both the proposed pixel-wise and region-wised based similarities give the best performance in terms of all quantitative measurements including Global Consistency Error (GCE), Boundary Displacement Error (BDE), Variation of Information (VOI) and Probabilistic Rand Index (PRI) among all algorithms tested.

While popular image representation for classification such as Bag-of-Visual words (BoV) [42], Fisher Vector [43], and their variants [45-47] aggregate local descriptors in images and utilize Gaussian mixture representation of a large feature vector for classification, the proposed probabilistic representation of color image pixels (PRCI) explores the mixture distribution in measuring the similarity of given pair of pixels and regions. Our primary aim is to explore the color distribution of an image or image sequence to develop a more discriminative measure to facilitate various operations such as image region matching, registration, etc. This is in sharp contrast to the mixture representation of features above in image classification, which usually involves high dimension features. On the other hand, PRCI alone may not be suitable for classification. However, it may be integrated with other features to form high-level features for classification.

The rest of this paper is organized as follows. In Section II, we review some related works on color naming and color descriptors. The proposed PRCI approach, including the clustering algorithm, membership evaluation in terms of both multivariate Gaussian and Laplace distributions, pixel-wise similarity and region-wise similarity, are described in Section III. Section IV is devoted to the application of the proposed pixel-wise PRCI similarity to a dense image descriptor-based multi-layered motion estimation problem. Another application, namely automatic image segmentation employing the proposed region-wise PRCI similarity is illustrated in Section V. Conclusions are drawn in Section VI.

[2] The approximation error will decrease with the number of representative regions used, which in turns is related to the number of clusters or representative vectors used in the clustering algorithm for approximating the color distribution of the input image(s). The Calinski-Harabasz index [6] is adopted as the criterion to determine the number of clusters, which was shown to yield good results in the computer experiments. More results on its performance can be found in the supplementary material.

The final version of record is available at http://dx.doi.org/10.1109/TIP.2018.2883580


## II. Related Works

In this section, we briefly review color names, a sub-category of "pure" color descriptors [39] and some other commonly used color descriptors. The term "pure" here refers to descriptors that utilize pixel-level information without local shape information. In addition, we also introduce some popular image-level probabilistic representations.

### A. Color Names

Color names models aim to mimic the usage of real-world color labels/vocabularies as a linguistic study [36] showed that eleven basic color terms, namely: black, blue, brown, grey, green, orange, pink, purple, red, white and yellow, are most commonly used in English language. Though there are different opinions towards the actual number of categories or partition methods [32, 33, 37], extensive studies have been carried out to link colors to pre-defined color categories. In [37], only four colors (red, green, yellow and blue) are used and they are the first to use fuzzy set theory to define membership of arbitrary color value to these four basic colors. Zheng et al. [32] extends the set of color names to 16 color names. Mojsilovic [34] first proposed to employ computational model for color categorization while Benavente et al. [35] formulated this problem in the framework of fuzzy set theory. Van de Weijer et al. [33] proposed to learn color names using real-world image and labels from Google image search engine and eBay product search engine with a Probabilistic Latent Semantic Analysis (PLSA) model instead of calibrated data. Khan et al. [38] extended van de Weijer's concept by refining the partition in Lab color space using Divisive Information-Theoretic Clustering (DITC) algorithm to optimize its discriminative power. These color categories can be viewed as a mixture of photometric invariant under various illumination conditions with high discriminating power across images at a very compact representation of only 11 dimensions.

### B. Other Color Descriptors and Image-level Probabilistic Representations

As mentioned above, color descriptors can be pure or regional. The pure ones usually involve projection of original 3-dimensional signals or image pixels onto space with larger dimensions while regional descriptors further include local geometrical cues. Khan et al. [39] summarized some commonly used pure color descriptors ranging from the simplest ones e.g. RGB descriptor, C Descriptor to higher level ones like robust hue descriptor and opponent derivative descriptor. On the other hand, regional color descriptors can be categorized into two types: 1) histogram-like descriptors such as Color correlogram [1], and Colored pattern appearance histogram [2]; and 2) regularized patch features-based descriptors such as Scale Invariant Feature Transform (SIFT) [3], and DAISY [4]. A comprehensive summary and comparison of their performances under object and scene recognition problems was given in [55].

Fisher Vector [43], Bag-of-Visual words (BoV) [42] and their variants [45-47] are popular probabilistic image representation for classification. These techniques start with local feature extraction such as histogram of gradients (HoG), SIFT [3] or Speed Up Robust Features (SURF) [44] descriptors. To aggregate these local descriptors into a global representation, the Fisher Vector (FV) models the local descriptors as a Gaussian Mixture Model (GMM) and measures the gradient of the sample log-likelihood function with respect to the model parameters as the image-level descriptor. The BoV, however, performs clustering on all local descriptors from training images to create a set of bases. It then pools image-wise local descriptors onto these bases for an image-level representation. Interested readers are referred to [56] for a detailed comparison of FV and BoV. Recently studies are focused on using better coding techniques such as soft assignment and sparse coding under these frameworks.

As mentioned, the proposed probabilistic color representation also models the color pixels as mixtures to construct various local or regional similarity aiming at improving the performance of image region matching, registration and related applications. Specifically, we propose to use the posteriori probabilities of a color pixel in each cluster, or its membership, as a feature or membership vector for constructing more general similarity measures. This has the advantage that fewer samples are needed to capture the local pdf by exploring the prior information of the mixture components. Furthermore, the posteriori probability for both multivariate Gaussian and multivariate Laplacian mixtures are derived. Unlike the conventional multivariate Gaussian mixture, the multivariate Laplacian mixtures has the advantage of reduced sensitivity to outlying samples and uncertainties, which leads to a more robust representation.

## III. The Proposed Probabilistic Representation of Color Image Pixels (PRCI) And Similarity Measures

In this section, we first focus on the determination of the mixture representation using an efficient vector quantization (VQ) algorithm to provide a reliable partitioning of the color space. This can be further refined using more sophisticated clustering algorithm such as the EM algorithm. We then propose in subsection III-B a Probabilistic Representation of Color Image Pixels (PRCI) in terms of their memberships in the component densities of the mixture. To reduce the sensitivity of the mixture to modeling errors, we further derive the membership for the multivariate Laplace component density. Based on this component representation of color pixels, we propose in subsection III-C a pixel-wise similarity measure of two given pixels by approximating the continuous domain Bhattacharyya coefficient. This yields a convenient expression in terms of their memberships in the finite component mixture. Furthermore, in subsection III-D, this pixel-wise similarity is extended to measure the similarity between two image regions.

### A. Determination of the Color Mixture

Like many probabilistic color descriptors [32-35, 37, 38], the goal of determining the mixture representation of the color pixels of an image or image sequences is to partition the pixels into representative color groups according to certain criterion. This is closely related to vector quantization (VQ) and clustering. In VQ [50], the image pixels are quantized (or clustered) to their nearest representative vectors so as to minimize the sum-of-squares errors. The representative vectors or centers of the clusters are then updated as the mean of the pixels associated with them. Because of the latter operations, it is also referred to as the K-means algorithm, which is frequently used in clustering. On the other hand, other criteria can be employed in the K-means algorithm to measure the distance between a given pixel to the cluster centers [7]. When clustering a pixel to the clusters, one can

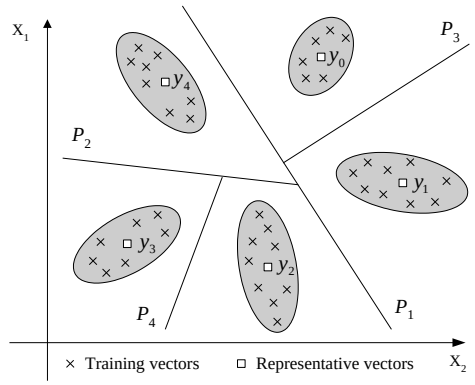


Figure 2. Illustration of the AHPVQ algorithm [5] for a set of two dimensional vectors $\boldsymbol{x}=[X_1, X_2]^T$. The representative vectors or clustering centers are denoted by $y_i$, $i$=0,…,4. $P_i$, $i$=1,…,4, denote the partitioning hyperplanes.

use the hard membership (also known as winner-take-all) approach as in conventional VQ where a pixel will be assigned to one and only one cluster. Alternatively, one can use the soft membership approach where the probability that the pixel is belonging to a given cluster is used to update the corresponding cluster mean.

If the pixels in each cluster are modeled as a multivariate Gaussian distribution, one can describe them conveniently with its mean and covariance matrix. Moreover, the mean and covariance matrix of each cluster $C_n$, $n$=1, …, $N$, can be estimated empirically from the samples in the corresponding cluster. Here, we assumed that there are $N$ clusters. The mean can be viewed as the representative color of the *n-th* cluster in a given color space, whereas the covariance represents the spread of the color pixels inside the given cluster. This can also be viewed as a Gaussian mixture (GM) approximation to the pdf of the original data set. In traditional VQs, each color pixel is represented by the mean of its closest cluster. This is a simplification of the pdf by retaining only the mean components. The GM approximation is also related to the local PCA (LPCA) representation proposed in [55] for machine learning where the covariance matrices may be further simplified by retaining the dominant eigenvalues and eigenvectors. Since the covariance matrix in our case is a (3x3) matrix, such simplification may not be necessary. We shall see later that not only does the mixture distribution yield a more precise description than the representative colors, it also allows us to devise many image descriptors from this simplified pdf. Since we are interested in incorporating the prior color mixture component distributions in deriving similarity and distance measure to facilitate various image processing problems, the algorithm for finding the component distributions should be reliable and fast. Also, it is desirable to be able to determine automatically the number of components or clusters.

To this end, we adopt an efficient Arbitrary HyperPlane Vector Quantization (AHPVQ) algorithm [5] to partition the image pixels into groups and estimate their corresponding mean and covariance matrices. AHPVQ is a hierarchical clustering algorithm where the original data is successively partitioned by a series of separating hyperplanes to obtain the desired clusters. This is illustrated in Fig.2 where the input data is partitioned into five clusters. The training data is first divided using a hyperplane $P_1$: $\boldsymbol{h}_1^T \boldsymbol{s} + \beta_1 = 0$ into two partitions corresponding to those vectors respectively on the left and right of the hyperplane where $\boldsymbol{s} \in R^3$ is the given color space, $\boldsymbol{h}_1 \in R^3$ is the normal vector of $P_1$, and $\beta_1$ is an appropriate constant. This partitioning can also be viewed as designing a codebook with two elements. Then, one of these partitions will be chosen according to a certain criterion for further subdivision. The process repeats at each stage by choosing an appropriate partition for further subdivision until the desired number of partitions (i.e. number of clusters) is obtained. The AHPVQ [5] aims to minimize the total sum-of-squared errors (SSE) by splitting the cluster which will lead to the largest reduction in SSE. To determine the hyperplane for splitting a given cluster into two clusters, the corresponding hyperplane, say $P_k$, is chosen to pass through the mean of the original cluster and its normal $\boldsymbol{h}_k$ is chosen as the largest eigenvector of the covariance matrix of the data inside the given cluster. This allows the direction with the largest spread to be partitioned so as to reduce the SSE. The hyperplane can be further refined by using any K-means algorithms. Due to the hierarchical nature and the use of the largest eigenvector of the associated data covariance matrix, the AHPVQ algorithm yields a very stable partitioning of the original data with good performance. This is in contrast to conventional K-means algorithms where multiple initializations have to be performed to find the best partitions in order to achieve a good SSE performance. In addition, due to its hierarchical nature, it can generate all partitions with cluster number less than or equal to the target number and the complexity in increasing the cluster number is very low. This is particularly useful when one is interested in finding an appropriate cluster number by examining their performances with increasing cluster numbers. Thus, it can be used directly to estimate the component densities or serving as a stable initial guess for further refinement using other clustering methods to reduce the computational requirement.

As an illustration, Fig.1(c) shows the clusters obtained by applying the AHPVQ algorithm to the image in Fig. 1 (a) where the sum-of-square errors is minimized by 20 clusters. It shows that the overall shape of the pdf is captured in the component representation. Regarding the selection of input color spaces, the RGB, Lab, YCbCr and HSV are frequently used in image processing algorithms. Among them, the RGB color space is the most popular as it is widely used in video analog-to-digital converters. Another popular option is Lab color space, where the distance used is more consistent with perceptual difference. Regarding the selection of number of clusters $N$, which is usually set empirically, we propose to employ clustering indices such as the Calinski-Harabasz index [6]. Its selection criteria are based on measuring the data partition by their within-class and between-class scatter.

Due to page limitation, the comparison of computational complexity and performance of using K-means and AHPVQ for the proposed PRCI method are included in the supplementary material. In subsequent experimentation, the AHPVQ algorithm with Lab color space is used due to its good performance and low arithmetic complexity. We also employ the Calinski-Harabasz index for automatic selection of the cluster number $N$.

### *B. Probabilistic Representation of Color Image Pixels (PRCI)*

After clustering the given image or image sequences, one gets a representation with say *N* major clusters. For instance, one can estimate the pdf of a group of pixels in terms of the mixture components. More precisely, let $\boldsymbol{s}_i \in R^3$ be the color of the *i-th* pixel in the dataset. We wish to determine the probability of $\boldsymbol{s}_i$ belonging to the *n-th* cluster, i.e. $p(C_n \mid \boldsymbol{s}_i)$. Using the Bayes' theorem, we can write $p(C_n \mid \boldsymbol{s}_i)$ in terms of the likelihood function of class *n*, $p(\boldsymbol{s}_i \mid C_n)$, and its prior probability $p(C_n)$ as follows

$$p(C_n \mid \boldsymbol{s}_i) = \frac{p(\boldsymbol{s}_i \mid C_n)p(C_n)}{p(\boldsymbol{s}_i)} = \frac{p(\boldsymbol{s}_i \mid C_n)p(C_n)}{\sum_{n=1}^{N} p(\boldsymbol{s}_i \mid C_n)p(C_n)}. \quad (1)$$

For simplicity, we assume that the prior probabilities $p(C_n)$, $n=1,\ldots,N$ are identical so that $p(C_n) = 1/N$. The generalization to given prior probabilities is straightforward. Therefore, equation (1) is reduced to

$$p(C_n \mid \boldsymbol{s}_i) = \frac{p(\boldsymbol{s}_i \mid C_n)}{\sum_{n=1}^{N} p(\boldsymbol{s}_i \mid C_n)}. \quad (2)$$

(2) suggests that the posterior probability $p(C_n \mid \boldsymbol{s}_i)$ under the uniform prior is actually the likelihood functions of class *n*, $p(\boldsymbol{s}_i \mid C_n)$, normalized by the sum of all likelihood functions in all the *N* classes. In other words, when the color of a given pixel is relatively close to one of the clusters, the corresponding posterior probability or membership for that cluster will be high. This suggests a vector-representation $\boldsymbol{m}(\boldsymbol{s}_i)$ of a color pixel $\boldsymbol{s}_i$ in a picture group in terms of the *N* normalized membership (2) in *N* independent cluster distributions.

$$\boldsymbol{m}(\boldsymbol{s}_i) = \left[p(C_1 \mid \boldsymbol{s}_i), p(C_2 \mid \boldsymbol{s}_i), \ldots, p(C_N \mid \boldsymbol{s}_i)\right]^T. \quad (3)$$

This membership representation of color pixels allows better discrimination between two color pixels in the image set due to the incorporation of the prior information of the color clusters. Motivated by this potential advantage, we aim to develop various probabilistic similarity measures based on this membership concept and investigate their application to image matching and registration problems.

Next, we turn to the modeling of the likelihood functions in (3). As mentioned, the clusters in several algorithms such as the expectation-maximization (EM)-based K-means algorithm [7] are assumed to be Gaussian distributed. The corresponding likelihood function $p(\boldsymbol{s}_i \mid C_n)$ is:

$$p_G(\boldsymbol{s}_i \mid C_n) \equiv \mid 2\pi\boldsymbol{\Sigma}_n \mid^{-1/2} \exp\left(-\tfrac{1}{2}(\boldsymbol{s}_i - \boldsymbol{\mu}_n)^T \boldsymbol{\Sigma}_n^{-1}(\boldsymbol{s}_i - \boldsymbol{\mu}_n)\right), \quad (4)$$

where $\mid 2\pi\boldsymbol{\Sigma}_n \mid$ denotes the determinant of the matrix $2\pi\boldsymbol{\Sigma}_n$. In conventional VQ, the winner-take-all update [3] using Euclidean distance $\| \boldsymbol{s}_i - \boldsymbol{\mu}_n \|_2$ or the following Mahalanobis distance [8] are frequently used.

$$d_M(\boldsymbol{s}_i, n) = \sqrt{(\boldsymbol{s}_i - \boldsymbol{\mu}_n)^T \Sigma_n^{-1}(\boldsymbol{s}_i - \boldsymbol{\mu}_n)}. \quad (5)$$

One can notice that $p(\boldsymbol{s}_i \mid C_n) = \mid 2\pi\boldsymbol{\Sigma}_n \mid^{-1/2} \exp(-\frac{1}{2} d_M^2(\boldsymbol{s}_i, n))$. Given the clustered samples, one can estimate empirically,

[3] That is, each vector is clustered to the cluster with the smallest criteria such as Euclidean or Mahalanobis distance.

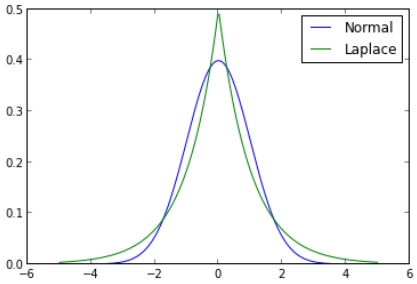


Figure 3. Comparison of univariate Normal distribution and Laplace distribution with same mean and variance.

the mean $\boldsymbol{\mu}_n$ and covariance $\boldsymbol{\Sigma}_n$ of the *n-th* cluster as $\hat{\boldsymbol{\mu}}_n = \Sigma_{\boldsymbol{s}_j \in C_n} \boldsymbol{s}_j$ and $\hat{\boldsymbol{\Sigma}}_n = \Sigma_{\boldsymbol{s}_j \in C_n} (\boldsymbol{s}_j - \hat{\boldsymbol{\mu}}_j)(\boldsymbol{s}_j - \hat{\boldsymbol{\mu}}_j)^T$, respectively. Consequently, the likelihood function in (4) can be evaluated and the resultant posterior probability is

$$p(C_n \mid \boldsymbol{s}_i) = \frac{\left|\Sigma_n\right|^{-1/2} e^{-d_M^2(\boldsymbol{s}_i, C_n)/2}}{\sum_{k=1}^{N} \left|\Sigma_k\right|^{-1/2} e^{-d_M^2(\boldsymbol{s}_i, C_k)/2}} \equiv p_G(C_n \mid \boldsymbol{s}_i). \quad (6)$$

Though the Gaussian distribution is widely used in image processing and other areas, it is more sensitive to outliers and large deviation in modeling errors. This can be visualized in Fig. 3 where the standard normal distribution and a Laplace distribution with the same mean and variance are plotted. It can be seen that the Gaussian distribution has lower peak at the origin but decays faster on the tail. On the other hand, the Laplace distribution has a positive excess kurtosis (leptokurtic) with fatter tails.

In the context of modeling the likelihood function in our membership vector, the Gaussian distribution may yield less noise tolerance to estimation errors than the Laplace distribution. In fact, our experiment on dense descriptor softmask construction to be presented later shows that using the Laplace distribution yields better performance than that of the Gaussian distribution. We now focus on the modeling of the posterior probability using the following *k*-dimensional asymmetric multivariate Laplace distribution [54]

$$f(\boldsymbol{x}) = \frac{2e^{\boldsymbol{x}^T \Sigma^{-1} \mu}}{(2\pi)^{d/2} \left|\Sigma\right|^{1/2}} \left( \frac{\boldsymbol{x}^T \Sigma^{-1} \boldsymbol{x}}{2 + \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu}} \right)^{\nu/2} \cdot K_\nu\left(\sqrt{(2 + \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu})(\boldsymbol{x}^T \Sigma^{-1} \boldsymbol{x})}\right), \quad (7)$$

where $\boldsymbol{\mu}$ is the mean, $\Sigma + \boldsymbol{\mu}^T \boldsymbol{\mu}$ is the covariance, $\nu = (2-k)/2$ and $K_\nu(.)$ is the modified Bessel function of the second kind [48]. When $\boldsymbol{\mu}$ is equal to zero, the distribution is symmetric. We shall model a sample $\boldsymbol{s}$ in the *n-th* cluster as

$$\boldsymbol{s} = \boldsymbol{\mu}_n + \boldsymbol{x}_n, \quad (8)$$

where $\boldsymbol{x}_n$ follows a three dimensional symmetric Laplace distribution given by:

$$f_n(\boldsymbol{x}_n) = \frac{2}{(2\pi)^{3/2} \left|\Sigma_n\right|^{1/2}} \left(\frac{\boldsymbol{x}_n^T \Sigma^{-1} \boldsymbol{x}_n}{2}\right)^{-1/4} K_\nu\left(\sqrt{2\boldsymbol{x}_n^T \Sigma^{-1} \boldsymbol{x}_n}\right). \quad (9)$$

where *v* is equal to $(2-3)/2 = -0.5$. Consequently,

The final version of record is available at http://dx.doi.org/10.1109/TIP.2018.2883580


$$p(\boldsymbol{s}\,|\,C_n) = f_n(\boldsymbol{s}-\boldsymbol{\mu}_n) = \frac{2}{(2\pi)^{3/2}|\Sigma_n|^{1/2}}\left(\frac{d_M^2(\boldsymbol{s},n)}{2}\right)^{-1/4} \cdot K_{-\frac{1}{2}}(\sqrt{2}d_M(\boldsymbol{s},n)). \quad (10)$$

We note that for $v$=-0.5, $K_{-\frac{1}{2}}(u)$ can be further simplified to $K_{-\frac{1}{2}}(u) \approx \sqrt{\frac{\pi}{2u}}e^{-u}$. Consequently, (10) can be expressed as:

$$p(\boldsymbol{s}_i\,|\,C_n) \approx \frac{e^{-\sqrt{2}d_M(\boldsymbol{s}_i,n)}}{2\pi|\Sigma_n|^{1/2}d_M(\boldsymbol{s}_i,n)} \equiv p_L(\boldsymbol{s}_i\,|\,C_n). \quad (11)$$

The corresponding posterior probability is then given by

$$p(C_n\,|\,\boldsymbol{s}_i) = \frac{(d_M(\boldsymbol{s}_i,C_k))^{-1}|\Sigma_n|^{-1/2}e^{-\sqrt{2}d_M(\boldsymbol{s}_i,C_n)}}{\sum_{k=1}^{N}(d_M(\boldsymbol{s}_i,C_k))^{-1}|\Sigma_k|^{-1/2}e^{-\sqrt{2}d_M(\boldsymbol{s}_i,C_k)}} \equiv p_L(C_n\,|\,\boldsymbol{s}_i). \quad (12)$$

Comparing (12) with (6), we found that the argument of the exponential function of the multivariate Gaussian distribution and Laplace distribution are proportional to $-d_M^2(\boldsymbol{s}_i,n)$ and $-d_M(\boldsymbol{s}_i,n)$, respectively. Moreover, the likelihood function of the multivariate Laplace distribution is inversely proportional to $d_M(\boldsymbol{s}_i,n)$, the Mahalanobis distance. Consequently, the posterior probability of multivariate Laplace distribution will approach infinity as the sample $\boldsymbol{s}_i$ approaches the cluster mean $\boldsymbol{\mu}_n$. This may not be desirable if one wishes to anticipate the effect of noise or uncertainty whereby a color pixel being close to a particular class may still have a small possibility of belonging to a neighboring class. Moreover, the inverse may also create numerical problem in fixed point implementation. Therefore, we propose to simplify (12) by removing the inverse of Mahalanobis distance as follows:

$$\tilde{p}_L(C_n\,|\,\boldsymbol{s}_i) = \frac{|\Sigma_n|^{-1/2}e^{-\sqrt{2}d_M(\boldsymbol{h},D_n)}}{\sum_{k=1}^{N}|\Sigma_k|^{-1/2}e^{-\sqrt{2}d_M(\boldsymbol{h},D_k)}}. \quad (13)$$

The membership vector with the simplified multivariate Laplace distribution for a given color pixel $\boldsymbol{s}_i$ is denoted by $\tilde{\boldsymbol{m}}(\boldsymbol{s}_i) = [\tilde{p}_L(C_1\,|\,\boldsymbol{s}_i), \tilde{p}_L(C_2\,|\,\boldsymbol{s}_i), \ldots, \tilde{p}_L(C_N\,|\,\boldsymbol{s}_i)]^T$. Due to page limitation, the comparison of membership vectors based on the multivariate Gaussian and Laplacian distributions on classic image registration using free-form deformation problem is given in the supplementary material. It is found that the Laplacian membership yields better performance than its Gaussian counterpart.

As mentioned earlier, the PRCI can better capture the local pdf as the use of the mixture components as the prior information can significantly simplify the estimation of the local probability density function (pdf) of an image region. To see this, since $m_n(\boldsymbol{s}_i) = p_L(C_n\,|\,\boldsymbol{s}_i)$, the *n-th* element of $\boldsymbol{m}(\boldsymbol{s}_i)$, is the posterior probability that $\boldsymbol{s}_i$ belongs to the *n-th* clusters and $p(\boldsymbol{s}\,|\,C_n)$ is the probability of a color $\boldsymbol{s}$ given that it is from the *n-th* mixture, the underlying pdf given the observed pixel $\boldsymbol{s}_i$ can be written as

$$p(\boldsymbol{s}\,|\,\boldsymbol{s}_i) \approx \sum_{n=1}^{N} p(\boldsymbol{s}\,|\,C_n)m_n(\boldsymbol{s}_i). \quad (14)$$

Here, we have assumed that pdf of the colour pixels can be represented as a linear combination of the mixture components which have been estimated from the image(s). We can see that the posterior probabilities serve as the mixing coefficients in the mixture for modeling the underlying pdf given the sample $\boldsymbol{s}_i$. Using this representation, the estimation of the local pdf of a group of pixels is reduced to the expectation of the membership vector over the samples over the image region. Moreover, the estimated pdf can be utilized to evaluate analytically or numerically the mathematical expectation of a given quantity. The concept is elaborated further in Section VI of the supplementary material, where the first and second statistics of an image region are expressed in terms of the membership of the samples. Next, we shall introduce a new pixel-wise similarity measure and distance based on the above PRCI mixture representation. It is then extended to a region-wise similarity, which will be utilized in various image processing applications.

### *C. PRCI-Based Similarity Measure and Distances*

#### *1)* Pixel-wise Similarity

In many applications that involve pixel level feature extraction, it is important to be able to measure the difference/similarity between the two colors. The proposed PRCI approach measures the similarity between two pixels $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ based on the Bhattacharyya coefficient (BC) [9] of their underlying estimated mixture pdf representation. The BC is an approximate measurement of the amount of overlap between two statistical samples and it gives the relative closeness of the two samples. The interval of the values of the two samples is split into a chosen number of partitions, say *N*, and the BC is determined according to the number of members of the two samples in the *n-th* partition, $p_n$ and $q_n$, as follows

$$BC(\boldsymbol{p},\boldsymbol{q}) = \sum_{n=1}^{N}\sqrt{p_n \cdot q_n}, \quad (15)$$

where $\boldsymbol{p} = [p_1,..., p_N]^T$ and $\boldsymbol{q} = [q_1,..., q_N]^T$. Obviously, the larger the overlap in the membership in the intervals, the larger will be the resulting BC. If there is no overlap, then the BC is zero, because either $p_n$ or $q_n$ will be zero. Since the PRCI vector of a pixel measures the membership of the color pixel for a particular representative color, it can be viewed a partitioning of the 3D color space similar to the Voronoi region in VQ. Thus, the BC can be used to measure the closeness of the two color distributions $p(\boldsymbol{s}\,|\,\boldsymbol{s}_i)$ and $p(\boldsymbol{s}\,|\,\boldsymbol{s}_j)$ associated with the two observations with $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$ by replacing $\boldsymbol{p}$ and $\boldsymbol{q}$ in (6) respectively by $\boldsymbol{m}(\boldsymbol{s}_i)$ and $\boldsymbol{m}(\boldsymbol{s}_j)$. More specifically, it is shown in the appendix that the BC of $p(\boldsymbol{s}\,|\,\boldsymbol{s}_i)$ and $p(\boldsymbol{s}\,|\,\boldsymbol{s}_j)$ can be written as

$$\psi(\boldsymbol{s}_i,\boldsymbol{s}_j) \approx \sum_{n=1}^{N}\sqrt{m_n(\boldsymbol{s}_i)m_n(\boldsymbol{s}_j)} = \sqrt{\boldsymbol{m}(\boldsymbol{s}_i)^T\boldsymbol{m}(\boldsymbol{s}_j)}. \quad (16)$$

Since $m_k(\boldsymbol{s}_i) \geq 0$ and $\Sigma_{n=1}^{N}m_n(\boldsymbol{s}_i) = 1$, we have $0 \leq \psi(\boldsymbol{s}_i,\boldsymbol{s}_j) \leq 1$. Based on the BC, we can also define various distances between $\boldsymbol{s}_i$ and $\boldsymbol{s}_j$. For instance, the Bhattacharyya distance (BD) [9] is defined as

$$D_B(\boldsymbol{s}_i,\boldsymbol{s}_j) = \ln(BC(p(\boldsymbol{s}\,|\,\boldsymbol{s}_i), p(\boldsymbol{s}\,|\,\boldsymbol{s}_j)) \quad (17)$$

$$\approx \ln(\sqrt{\boldsymbol{m}(\boldsymbol{s}_i)^T \boldsymbol{m}(\boldsymbol{s}_j)}).$$

The BD satisfies $0 \le D_B(\boldsymbol{s}_i, \boldsymbol{s}_j) \le \infty$ but it does not obey the triangular inequality. On the other hand, the Hellinger distance [49]

$$D_H(\boldsymbol{s}_i, \boldsymbol{s}_j) = \sqrt{1 - BC(p(\boldsymbol{s} \mid \boldsymbol{s}_i), p(\boldsymbol{s} \mid \boldsymbol{s}_j))} \approx \sqrt{1 - \sqrt{\boldsymbol{m}(\boldsymbol{s}_i)^T \boldsymbol{m}(\boldsymbol{s}_j)}} \quad (18)$$

does obey the triangular inequality. Therefore, we may also use these distances instead of the traditional Euclidean distance $\| \boldsymbol{s}_i - \boldsymbol{s}_j \|_2^2$ in various applications to reflect the prior information of the color clusters.

*2)* Region-wise Similarity

In contrast to pixel-wise similarity, the precise definition of a good region-wise similarity is less straightforward, since the similarity measures are usually based on human intuition. To measure color-only similarity, we can use the first/second order statistics of regional color distribution or compare the overlap in the color histograms. To take structural similarity into consideration, one may employ spatial heterogeneity statistics such as *q-stat* [10] to measure the significance level of heterogeneity between two regions. For more robust higher-level regional descriptors, one can resort to uniformly sampling techniques such as SIFT, Daisy, etc. In this subsection, we focus on analyzing the color-based similarity between the regional color distributions and propose a new region-wise similarity based on the PRCI framework. In particular, we shall focus on a PRCI-based histogram.

The conventional histogram of a group of image pixels is the frequency density of color values in a set of pre-defined regions, usually obtained by divided along each of the color space dimensions. It represents the distribution of color in an image or region of interest while ignoring the spatial location of the colors. The local histogram can provide useful information in discriminating different image regions and hence objects and textures. However, different selections of color space may yield different characteristics. Moreover, a small bin size is usually chosen in color histogram to give a more accurate statistical characterization.

Since the PRCI is based on the partitioning of the color space into $N$ regions, it is also quite convenient to calculate the color histogram of a given image region. More specifically, we can simply divide or quantize the range of each membership or mixing coefficient into bins as in conventional histogram. By quantizing the membership coefficients of the given image region $R$ to the nearest bins, one obtains the frequency of each bin, and hence $N$ histograms each coming from one of the $N$ membership coefficients. Let the normalized histograms of the $n$-th membership coefficient of $R$ be $H_{n,b}(R)$, where $b = 1,\ldots,B$, denotes the bin index and $B$ is the number of bins. Because of the normalization, we have $\Sigma_{b=1}^{B} H_{n,b}(R) = 1$ $0 \le H_{n,b}(R) \le 1$. For simplicity, we assume that the number of bins is the same for all membership coefficients. They can also be chosen to be different numbers if necessary.

Since histogram is a quantized pdf, it is also possible to measure the similarity of two different image regions $R$ and $R'$ with sets of pixels $\{\boldsymbol{s}_k \mid k = 1,\ldots,K\}$ and $\{\boldsymbol{v}_{k'} \mid k' = 1,\ldots,K'\}$ respectively. More precisely, let $H_{n,b}(R)$ and $H_{n,b}(R')$, $b = 1,\ldots,B$, be respectively the histograms of the $n$-th membership coefficients for $R$ and $R'$. The similarity between $H_{n,b}(R)$ and $H_{n,b}(R')$ for each $n$ can be approximated by the BC as follows

$$\Phi_n(R, R') = \sum_{b=1}^{B} \sqrt{H_{n,b}(R) H_{n,b}(R')}, \; n = 1,\ldots,N. \quad (19)$$

Therefore, $0 \le \Phi_n(R, R') \le 1$. The similarity taking all $n$ from 1 to $N$ into account can then be defined as the average of $\Phi_n(R, R')$ over all $n$ as follows:

$$\Phi(R, R') = \frac{1}{N} \sum_{n=1}^{N} \Phi_n(R, R'). \quad (20)$$

Again, $0 \le \Phi(R, R') \le 1$. Therefore, $\Phi(R, R')$ can serve as a measure of region-wise similarity, which incorporates the prior color distribution of the image or image set. The corresponding BD and Hellinger distance can be defined as $D(R, R') = \frac{1}{N} \sum_{n=1}^{N} D_n(R, R')$, where $D_n(R, R')$ can be chosen as $D_{B,n} = \ln \Phi_n(R, R')$, and $D_{H,n} = \sqrt{1 - \Phi_n(R, R')}$, respectively.

In the following sections, we shall focus on two practical image processing applications in order to illustrate the usefulness of the proposed PRCI-based approach over traditional color space representation.

## IV. Applications of Pixel-wise PRCI Similarity

In this section, we evaluate our pixel-wise PRCI similarity in dense descriptor soft segmentation mask (Softmask) construction and compare its performance with state-of-the-art softmask. Specifically, we build a soft-segmentation mask (a.k.a. "Softmask") using the proposed PRCI color similarity and employ it in a scale/rotation invariant dense (SID) image descriptor. We compare our descriptor with the state-of-the-art methods and the results show that the proposed method gives substantial improvement in multi-layer motion estimation.

### *A. PRCI-based Dense Image Descriptor*

Dense image descriptors such as dense SIFT [11], dense SID [12] and Daisy [4] have been proven to be effective in Bag-of-Words classification systems [42] as well as stereo estimation. Much focus has been given to the handling of non-rigid deformations, scale and occlusions. Regarding occlusion handling, which involves suppression of background information from foreground objects and removal of irrelevant features, soft segmentation is usually employed as a weighting mask in these descriptors [13-15]. Traditional soft segmentation handling occlusion include predefined set of binary mask [4], contour generated cue [13], and Normalized Cut eigenvectors [15]. In [16], a soft segmentation mask was incorporated in the state-of-the-art dense Scale-and Rotation-Invariant descriptor (SID) [12], a dense intensity-based descriptor. Substantial performance improvement was achieved in large-displacement, multi-layered motion estimation and wide-baseline stereo. Motivated by the potential advantages of the proposed pixel-wise PRCI similarity, a new PRCI-based "Softmask" is proposed below as an alternative to the Local Color Model (LCM) "Softmask" used in [12]. Experimental results show

that substantial improvement can be achieved in multi-layer motion estimation.

*1) Descriptor Construction*

As mentioned earlier, a soft segmentation weighting mask can be embedded in most state-of-the-art descriptors such as dense SIFT, Daisy and dense SID to suppress irrelevant information which is geometrically remote from the feature center. Usually, pixels with low weighting are considered topologically far away from the center pixel. Here, we focus on SID as it can achieve scale- and rotation- invariance which is suitable for object tracking and easy to implement. The SID descriptor starts with a log-polar sampling and compute gradient features around every sample position. The log-polar grid associated with a feature center consists of $M$ rings and $K$ rays at angles $\theta_k = 2\pi k/K$. The image derivatives at each sample position are calculated at 4 evenly distributed orientations and two polarities, with one pair of orientations steered to align with the angle directions. Such sampling scheme creates a $K-\text{by}-M$ matrix $\boldsymbol{D}$ which contains the corresponding sampled gradients given each orientation-polarity combination. Each row of the matrix represents the samples on the same ray from all rings, whereas each column represents the samples on the same ring from all rays. Given a soft segmentation mask $w_{i,j} \in [0,1]$, it is multiplied to the measurements $\boldsymbol{D}$ at each sampling grid point as:

$$\overline{D}_{i,j} = w_{i,j} D_{i,j}, \quad (i,j) \in [1,K] \times [1,M] \tag{21}$$

where $D_{i,j}$ is the $(i,j)$-*th* element of matrix $\boldsymbol{D}$.

Applying the weights $w_{i,j}$ to the descriptor effectively shuns measurements coming from irrelevant areas such as occlusions, background, and other objects. Therefore, the features extracted from this center point will remain robust to background changes. The formation of traditional soft segmentation weighting mask and the proposed PRCI-based "Softmask" will be discussed in the next subsections. After calculating the weighted descriptor $\overline{D}_{(i,j)}$, Fourier transform is then applied along every row to obtain the magnitudes of the Fourier coefficients. Thus, samples at different distances are now turned into frequency domain and the resulting descriptor becomes scale-invariant. If Fourier transform is also applied to each column, the output will be both scale- and rotation-invariant.

*2) Soft Segmentation*

The goal of building a soft segmentation mask is to effectively suppress irrelevant information caused by either changes of background when examining correlation of two locations. As it is difficult to recover the missing information, removing such outlying observations will still improve the robustness of a feature-based appearance descriptor. The major problem in constructing a soft segmentation mask is to assign a similarity measure between two given points or regions which yields a small distance measure only when the two points are within the same object. Since exact segmentation is still an open question to be tackled, a useful approach is to differentiate local regions into multiple labels with different similarity in the form of a "Softmask". In [13], a boundary cue probability $Pb_{\sigma}(x,y,\theta)$ which integrates multiple cues to estimate a boundary-based affinity using the "intervening contour" technique of [17] was proposed. These local affinities are subsequently 'globalized' by finding the eigenvectors of the relaxed Normalized Cut criterion. Leordeanu et al., [14] on the other hand, proposed to use a Local Color Model (LCM) within the neighborhood of a pixel to construct a large set of figure segmentations. It is then projected to a lower dimensional subspace using principal component analysis. These methods are more discriminative compared with the naive Euclidean distance in color space as they possess considerable additional information at the expense of increased computational complexity.

In our proposed approach, instead of building complex color models, we propose to integrate our pixel-wise PRCI to this framework via soft segmentations. Comparing with simple Euclidean distance, the similarity between pixels in terms of the underlying mixture component distributions is a stronger cue to separate them. The mixture components can further shun the influence of occluders or background noises as they are usually classified into different color components. As for the similar colors in local area, their similarity can be measured more accurately in terms of the component distribution. Comparing with the aforementioned "Softmasks" in [13, 14], our approach focuses on increasing the discrimating power via the global color distribution, which can also be served as an addendum to the boundary cue probability in [13].

*3) Experimental Setup*

We examine our PRCI-based Softmask under the SID framework on multi-layer motion estimation and compare it to the state-of-the-art SSID (along with SSID-Rot version) [16]. For notation convenience, we denote the proposed PRCI-embedded Scale Invariant Descriptor by PRCI-SID. We follow the same setting of the SID sampling scheme in [12] with $M = 28$ rings, $K = 32$ rays and $D = 8$ orientations (4 orientations and both polarities). As the matrix is real, the magnitudes of the Fourier spectrum are symmetric so that only data in two quadrants are retained. The DC component is also discarded as it might be affected by lighting changes. The descriptor is normalized to have unit $L_2$ norm with a size of 3328 for SID (SSID) and 3360 for SID-Rot (SSID-Rot) respectively.

The multi-layered motion tracking experiment is performed on the Berkeley Motion Dataset (Moseg) [18] which contains 10 video sequences of outdoor traffic taken by a handheld motion camera, 3 sequences of moving people, as well as 13 sequences from TV series "*Miss Marple"*. All these sequences exhibit multi-layered motion. Ground truth is provided by the dataset about every ten consecutive frames. We performed our tests on 5 traffic sequences, 3 "*Miss Marple"* sequences and 2 moving people sequences by going through all possible pairings of frames with ground truth, totaling 482 frame pairs (224/140/80/38 pairs for 10+/20+/30+/40+ frame difference). Due to the high memory requirement of dense descriptors (3328×640×480×4 (single-precision floating-point) ×2 (two images) ≈ 8*GB*), every image is resized to 25% of its original size to limit the memory requirement and achieve a reasonable computation time as in [16]. SIFT-Flow [19], a publicly available variant of optical flow method, is adopted to cope with the dense SIFT-like descriptor instead of raw pixels to estimate the correspondence between image pairs. The following performance metric is used to evaluate the correspondence

The final version of record is available at http://dx.doi.org/10.1109/TIP.2018.2883580


accuracy using different descriptors: 1) the PSNR between the warped reference frame and the target frame and 2) Dice coefficient [20] which computes the overlap between the warped segmentation mask of the reference frame to that of the target frame.

### *4) Experimental Results*

Fig. 4 shows the PSNR and dice coefficient results of the proposed PRCI-SID under various mixture model settings and the SSID as well as its rotation-invariant version. The eigenvector embedding method [13] is not included as it gives inferior performance than LCM embedding as demonstrated in [16]. We observe that our PRCI embedding under all three mixture component settings consistently outperforms the state-of-the-art SSID and its rotation invariant version. From the PSNR measurement in fig. 4 (a), substantial improvement in PSNR (~3dB) for our PRCI-SID over the SSID is observed. It suggests that with the PRCI-based "Softmask", the SID is capable of more accurate matching of image correspondence and reconstructing the target frames. For the dice coefficient in fig. 4 (b), the proposed PRCI-SID shows better performance with SSID in the first bin and the gap between PRCI-SID and SSID increases as the frame difference increases. This suggests that the proposed pixel-wise PRCI similarity is better than the aforementioned "Softmasks" in terms of consistently finding the relevant parts of the object and suppressing the background noise/occluders. Visual assessment as illustrated in fig. 5 also confirms the PSNR improvement as the warped reference image obtained by the proposed PRCI-SID resembles more closely the original target than the SSID. It can also be seen from the estimated motion in fig. 5 that the PRCI-SID yields better results at object boundaries and discontinuous areas. Due to the page limit, more results and comparisons on computational time are presented in the supplemantary material.

## V. Applications of Region-wise PRCI Similarity

In this section, we propose an unsupervised segmentation method using superpixels with region-wise PRCI similarity.

### *A. PRCI-based Unsupervised Image Segmentation*

Image segmentation is a classical image processing problem and it plays a key role in many high-level computer vision tasks. Graph-oriented methods, such as unsupervised spectral segmentation algorithms, were widely studied [21, 22, 24, 25, 51-53]. To save arithmetic complexity, an initial over-segmentation is first performed to group similar image pixels together. A graph with nodes representing a set of similar regions from over-segmentation and edges representing the affinity between nodes based on certain similarity measures is then formed. Finally, graph partitioning algorithm such as Transfer Cuts [21] and Ncuts [15] can be used to determine

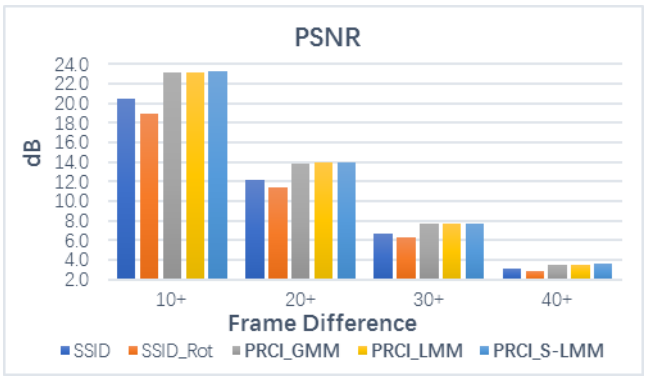


Figure 4. (a) PSNR results over the Moseg Dataset for PRCI-SID under different mixture models. The proposed methods are marked in bold. The results are accumulated so the first bin (10+) includes all frame pairs, the second bind (20+) contains pairs with a displacement of 20 or more frames, etc. Each bin shows the average overlap between all the frame pairs under consideration in terms of Dice Coefficient. The following figure follows the same protocol.

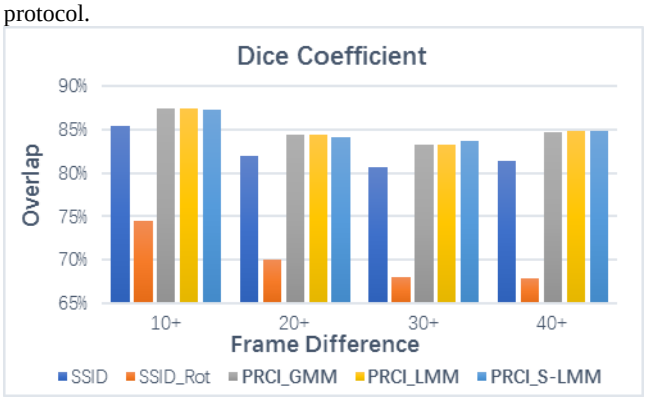


Figure 4. (b) Dice coefficient results over the Moseg Dataset under the same protocol of (a).

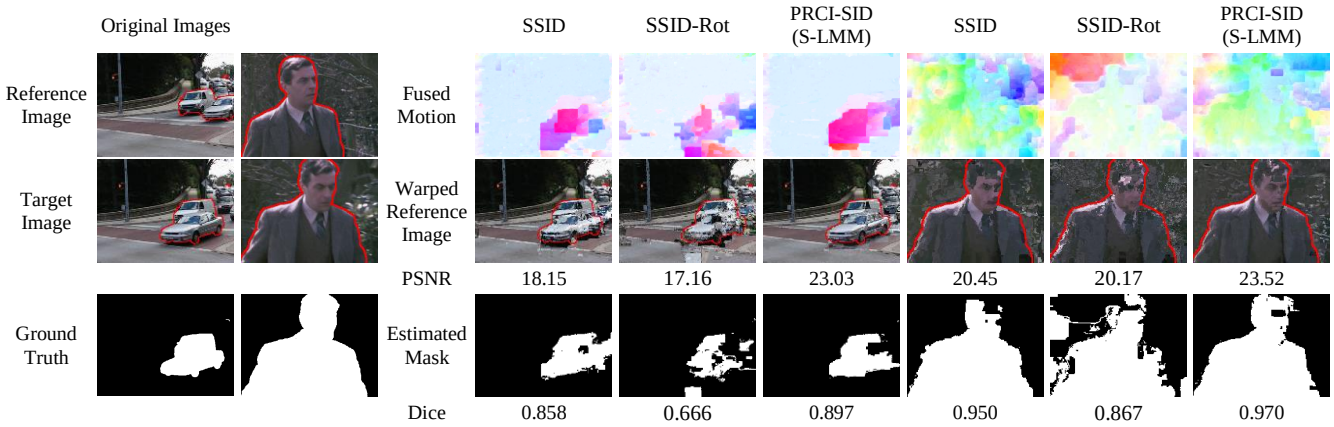


Figure 5. Results of large displacement multi-layer motion estimation using SIFT-flow. The reference image is registered to the target image with PRCI-SID, SSID and its rotation invariant mode. We warped the reference image to the target image using estimated correspondence and overlay the ground truth segmentation of target image in red to the warped reference image (a good registration should bring the object in alignment with the segmentation mask boundary in red). Fused motion (include both vertical and horizontal) are shown in color image of first row. We observe significant improvement in registration accuracy (much less noise and clear object boundary) as well as better visual similarity of warped image compared with target image. (Zoom in for more details)

the final segmentation.

As the quality of the final result depends highly on the initial segmentation, a number of methods have been proposed to determine reliably a convincing graph initialization using superpixels [21, 22]. These methods start with either a single or multiple over-segmentation (superpixel) methods under various parameter settings. On top of the initial segmentation, a graph was constructed by connecting the adjacent superpixels. Li et al. [21] constructed a bipartite affinity graph by assessing the similarity of neighboring superpixels for each superpixel layer and then solved for a global cut for all layers by Ncuts [15] or Transfer Cuts [21], namely Segmentation by Aggregating Superpixels (SAS). In [24], Wang et al. showed that a $l_0$ sparse representation in constructing an affinity graph can also achieve similar segmentation performance while reducing computational cost in building a bipartite graph. They also revealed in [25] a multi-scale version of [21] with better local/global homogeneity at the expenses of increased complexity. The multi-scale/criterion superpixel-based initialization has demonstrated their usefulness in capturing complex image structure. However, regarding the construction of the similarity measure between superpixels, there has no common agreement. In [22], Wang et al. investigated some popular similarity measures ranging from the simplest Euclidean distance between mean colors in Lab color space of superpixel (mLab) to higher level features such as Local Binary Pattern (LBP) [23] and SIFT. They also showed that the performance of first order color statistics or color histogram can be further improved.

In this work, we propose an alternative approach using our region-wise PRCI similarity as a new color cue for constructing the graph affinity. Comparing with mLab or color histogram (CH), our PRCI regional similarity explores the regional color distribution and is more resilient to illumination change such as shading, over-exposure, etc. It is expected to provide better discriminating power and robustness in the superpixel-based similarity measurement, which will be validated by the experimental results to be described below. In the next subsection, we briefly introduce our proposed regional PRCI similarity in the SAS framework.

*1) Regional PRCI similarity in SAS Framework*

In [21], a bipartite graph partition-based segmentation algorithm using multi-layer superpixels was proposed and the problem is solved via spectral clustering. The detail of bipartite graph formulation and transfer cuts, a spectral clustering method, can be found in [21].

**Table 1**. Performance evaluation of the proposed method (PRCI-S) against other methods over the Berkeley Segmentation Database.(Pix.: pixel-wise, Reg.: region-wise)

| Methods | | PRI ↑ | VoI ↓ | GCE ↓ | BDE ↓ |
|---|---|---|---|---|---|
| *GL-graph* [22] | *mLab* | 0.8230 | 2.0848 | 0.2260 | 11.71 |
| | *CH* | 0.8266 | 1.9585 | 0.2204 | 12.00 |
| *KNN-graph* [51] | *mLab* | 0.8290 | 2.0732 | 0.2316 | 12.19 |
| | *CH* | 0.8016 | 2.7882 | 0.3229 | 14.42 |
| *LRR-graph* [52] | *mLab* | 0.8155 | 1.8788 | 0.2071 | 13.70 |
| | *CH* | 0.8153 | 1.8794 | 0.2068 | 13.69 |
| $\ell_0$ *-graph* [24] | *mLab* | 0.8141 | 2.2969 | 0.2470 | 12.26 |
| | *CH* | 0.8185 | 2.2426 | 0.2622 | 12.84 |
| $\ell_1$ *-graph* [53] | *mLab* | 0.8036 | 2.9053 | 0.3079 | 12.77 |
| | *CH* | 0.7710 | 2.8919 | 0.3012 | 13.59 |
| *SAS* [21] | *mLab* | 0.8264 | 1.7537 | 0.1935 | 12.80 |
| | *CH* | 0.8133 | 1.9811 | 0.2204 | 13.96 |
| **PRCI-S (Proposed)** | Pix. | **0.8357** | **1.6732** | **0.1795** | **11.60** |
| | Reg. | **0.8364** | **1.6671** | **0.1799** | **11.25** |

In SAS, the affinity between nodes $i$ and $j$ (adjacent superpixels) is defined as $b_{ij} = e^{-\beta d_{ij}}$, where $d_{ij}$ is the Euclidean distance between the mean colors of two superpixels in Lab color space and $\beta$ is a constant weight similar to that in [21]. Our pixel-wise and region-wise PRCI similarities can be readily incorporated in the SAS framework by replacing $d_{ij}$ in the affinity $b_{ij}$ above with $D_H(\boldsymbol{s}_i, \boldsymbol{s}_j)$ in (18). For the region-wise similarity between two superpixels $R_i$ and $R_j$, the Hellinger distance $D_H(R_i, R_j) = \frac{1}{N}\sum_{n=1}^{N} D_{H,n}(R_i, R_j)$, where $D_{H,n} = \sqrt{1-\Phi_n(R_i, R_j)}$, mentioned in Section III-C is employed. Then, the graph affinity with region-wise PRCI becomes $b_{ij}' = e^{-\beta' D_H(R_i, R_j)}$, where $\beta'$ is a constant weight for $D_H$. Similar modification can be done in other state-of-the-art frameworks like $l_0$ graph [25] or G/L graph [22]. We now present the experimental results.

*2) Experimental Evaluation*

The performances of both pixel-/region-wise affinity measures $b_{ij} = e^{-\beta D_H(\boldsymbol{s}_i, \boldsymbol{s}_j)}$ and $b_{ij}' = e^{-\beta' D_H(R_i, R_j)}$ are evaluated on both the Microsoft Research Cambridge (MSRC) database

**Table 2.** Results of unsupervised segmentation on MSRC database. PRCI-S is SAS with our PRCI-based region-wise similarity, SAS is the original benchmark algorithm for comparison. Four quantitative methods are provided for each sub-categories and the superior one is colored and mark with bold and Italic.

| MSRC | PRI ↑ | | GCE↓ | | VoI↓ | | BDE↓ | | | PRI ↑ | | GCE↓ | | VoI↓ | | BDE↓ | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| *Catalog* | SAS | PRCI-S | SAS | PRCI-S | SAS | PRCI-S | SAS | PRCI-S | *Catalog* | SAS | PRCI-S | SAS | PRCI-S | SAS | PRCI-S | SAS | PRCI-S |
| ***1*** | 0.8866 | ***0.8920*** | *0.0887* | 0.0991 | *0.642* | 0.647 | *5.61* | 6.02 | ***11*** | 0.7709 | ***0.7756*** | *0.1277* | 0.1559 | *0.935* | 0.959 | 16.84 | ***13.53*** |
| ***2*** | 0.7346 | ***0.7457*** | 0.2753 | ***0.2671*** | 1.630 | ***1.611*** | 18.05 | ***16.11*** | ***12*** | 0.8161 | ***0.8324*** | 0.1138 | ***0.1054*** | 0.844 | ***0.798*** | 12.11 | ***9.77*** |
| ***3*** | 0.7765 | ***0.8025*** | 0.2528 | ***0.2434*** | 1.586 | ***1.522*** | 14.48 | ***10.33*** | ***13*** | 0.5502 | ***0.5619*** | *0.2621* | 0.2974 | *1.472* | 1.537 | 54.33 | ***43.31*** |
| ***4*** | 0.8789 | ***0.9052*** | *0.1670* | 0.1720 | 1.167 | ***1.109*** | 7.19 | ***5.80*** | ***14*** | 0.7228 | ***0.7549*** | 0.2379 | ***0.2344*** | 1.407 | ***1.344*** | 14.13 | ***12.53*** |
| ***5*** | 0.7630 | ***0.7969*** | 0.1897 | ***0.1592*** | 1.101 | ***0.959*** | 11.43 | ***8.98*** | ***15*** | 0.6338 | ***0.6572*** | *0.2337* | 0.2400 | 1.275 | ***1.241*** | 32.16 | ***27.56*** |
| ***6*** | 0.6365 | ***0.6643*** | *0.3362* | 0.3527 | *1.937* | 1.966 | 20.54 | ***18.41*** | ***16*** | 0.6866 | ***0.6920*** | *0.2228* | 0.2348 | *1.394* | 1.415 | 21.64 | ***20.23*** |
| ***7*** | 0.7117 | ***0.7290*** | 0.3103 | ***0.3007*** | 1.733 | ***1.687*** | 19.64 | ***18.02*** | ***17*** | 0.8142 | ***0.8256*** | 0.2566 | ***0.2530*** | 1.580 | ***1.514*** | 11.88 | ***11.55*** |
| ***8*** | 0.6401 | ***0.6677*** | 0.3810 | ***0.3728*** | 2.074 | ***2.045*** | 21.80 | ***18.90*** | ***18*** | *0.7794* | 0.7747 | *0.1634* | 0.1702 | *1.081* | 1.115 | *19.51* | 19.88 |
| ***9*** | 0.7739 | ***0.8047*** | 0.1454 | ***0.1305*** | 0.916 | ***0.828*** | 13.58 | ***11.89*** | ***19*** | 0.7690 | ***0.7925*** | 0.2734 | ***0.2608*** | 1.840 | ***1.804*** | 13.93 | ***11.07*** |
| ***10*** | 0.6679 | ***0.6800*** | *0.2483* | 0.2488 | 1.355 | ***1.339*** | 21.74 | ***17.96*** | ***20*** | *0.8715* | 0.8700 | *0.1630* | 0.1689 | *1.147* | 1.179 | 9.89 | ***9.38*** |

| *Average* | PRI ↑ | GCE↓ | VoI↓ | BDE↓ | | PRI ↑ | GCE↓ | VoI↓ | BDE↓ |
|---|---|---|---|---|---|---|---|---|---|
| *SAS [21]* | 0.7436 | 0.2226 | 1.356 | 17.98 | **PRCI-S** | **0.7608** | 0.2233 | **1.331** | **15.50** |

The final version of record is available at http://dx.doi.org/10.1109/TIP.2018.2883580


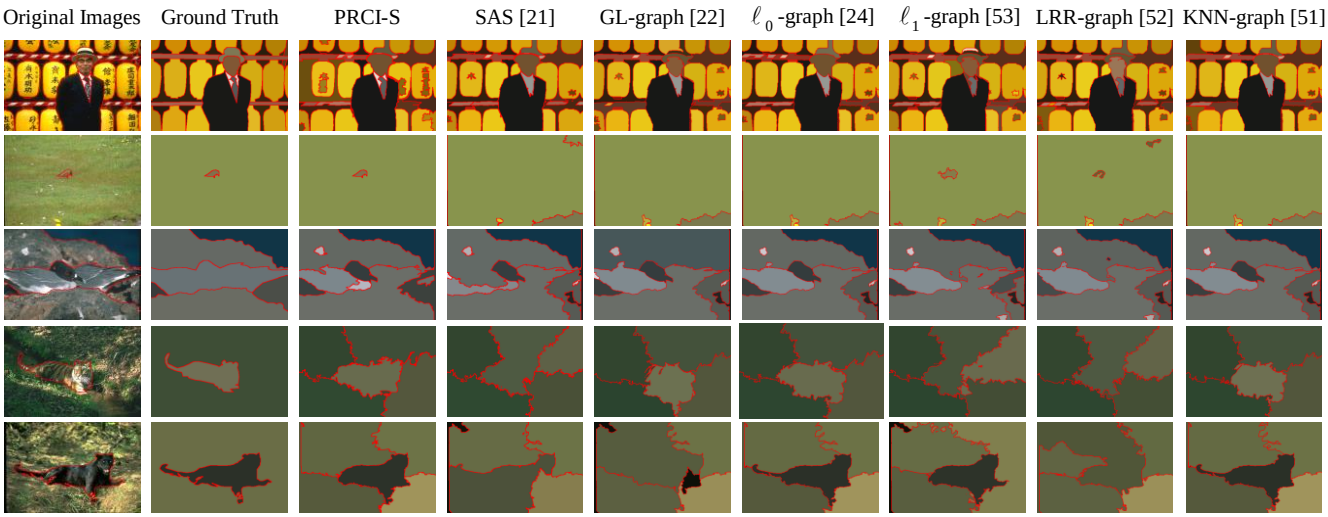


Figure 6. Visual comparison examples of our PRCI-S method and SAS method in BSD300 dataset. All methods adopt the same segmentation number *k* provided by [28].

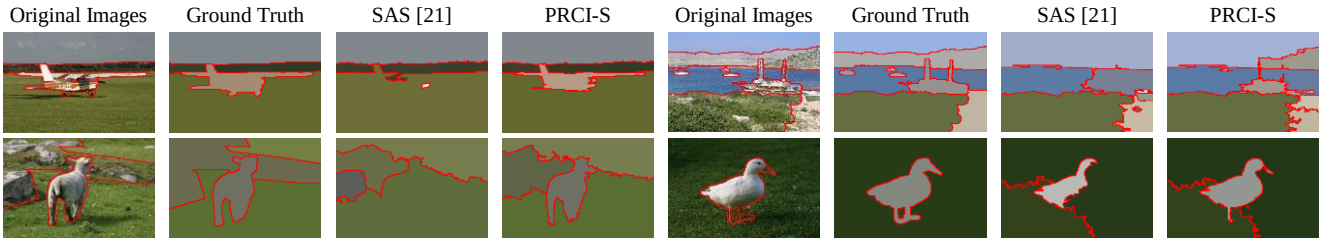


Figure 7. Visual comparison of our PRCI-S method and benchmark SAS method in MSRC dataset. Objects boundaries are marked in red.

[26] and BSD300 database [28]. The MSRC database contains 591 images from 23 object classes with ground-truth segmentation provided by [27]. The classified categories help us to verify the effectiveness of our algorithm under various circumstances by comparing with the benchmark method (SAS). For the BSD300 database [28], it consists of 300 natural images with multiple (~5) manual draw ground truth per image. To quantitatively evaluate the performance, we employ the following four popular performance metrics: 1) Probabilistic Rand Index (PRI) [29]; 2) Variation of Information (VOI) [30]; 3) Global Consistency Error (GCE) [28]; and 4) Boundary Displacement Error (BDE) [31]. Generally, segmentation is better if PRI is higher and the other three are smaller.

For the graph construction and partitioning, we proceed as in [21] where the over-segmented images are generated by the Mean Shift (MS) method and Felzenszwalb- Huttenlocher (FH) method with totally 5 or 6 settings. As for the similarity measurement, for fair comparison, we include both the mean Lab color space Euclidean distance (mLab) and color histogram (CH) similarity as affinity between graph nodes. We also include the results of other well established graph-based automatic segmentaion methods including GL-graph [22], KNN-graph [51], LRR-graph [52], $\ell_0$*-graph* [24] and $\ell_1$*-graph* [53] for comparison. The results of these methods are available in [22]. For our PRCI setting, we use Lab color space with automatic cluster number selection using Calinski-Harabasz index. The simplified-Laplcaian mixture model is also employed. More results and comparisons can also be found in the supplemantary material.

Table 1 shows the results of both our pixel-wise and region-wise similarity measurements in the SAS framework against other state-of-the-art methods on BSD300 dataset. The results show that our representation yields better performance in terms of all measurements, which indicates that our proposed PRCI-Similarity (PRCI-S) methods is a good alternative to other pure color features due to its robustness in gradual shading and discriminating power in color components. Table 2 summarizes the catalog-average scores of using the proposed PRCI-S and the benchmark Euclidean-based measure in the SAS for the MSRC database. PRCI-S outperforms the benchmark in most catalogs. In terms of overall performance across the whole dataset, our PRCI-S outperforms the mLab in PRI, VoI and BDE. Examples of our PRCI-S segmentation against state-of-the-art methods are shown in Figs. 6 and 7 for visual comparison.

## VI. Conclusion

A novel probabilistic representation of color image pixels (PRCI) and its application to derive pixel-wise and region-wise similarities for several image processing applications are presented. The usefulness of the proposed pixel-wise and region-wise similarities is demonstrated respectively in a dense image descriptor-based multi-layer motion estimation problem and an unsupervised image segmentation problem. Experimental results show that the proposed similarity yields improved PSNR performance and higher tracking accuracy in terms of Dice Coefficient over the state-of-the-art dense Scale-and Rotation-Invariant descriptor (SID) in multi-layered motion estimation. For the unsupervised image segmentation problem, both the proposed pixel-wise and region-wised based similarities give the best performance in terms of all quantitative measurements including Global Consistency Error (GCE), Boundary Displacement Error (BDE) Variation of Information (VoI) and Probabilistic Rand Index (PRI) among all algorithms tested. Future directions will focus on the generalization of the PRCI to include local structural information.

The final version of record is available at http://dx.doi.org/10.1109/TIP.2018.2883580


## APPENDIX – DERIVATION OF PRCI PIXEL-WISE SIMILARITY

Let the 3-dimensional space be partitioned according to the clustering centers into $N$ Voronoi regions as follows $V_n = \{\boldsymbol{s} \in R^3 \mid d_M(\boldsymbol{s}, C_n) \le d_M(\boldsymbol{s}, C_{n'}) \quad \forall\, n' \neq n\}$ , $n=1,\ldots,N$. The BC of $p(\boldsymbol{s} \mid \boldsymbol{s}_i)$ and $p(\boldsymbol{s} \mid \boldsymbol{s}_j)$ can be written as

$$\begin{aligned}\psi(\boldsymbol{s}_i, \boldsymbol{s}_j) &= BC(p(\boldsymbol{s} \mid \boldsymbol{s}_i), p(\boldsymbol{s} \mid \boldsymbol{s}_j)) \\ &= \int \sqrt{p(\boldsymbol{s} \mid \boldsymbol{s}_i) p(\boldsymbol{s} \mid \boldsymbol{s}_j)}\, d\boldsymbol{s} \,. \end{aligned} \tag{A.1}$$

Using (14), (A.1) can be written as

$$\begin{aligned}\psi(\boldsymbol{s}_i, \boldsymbol{s}_j) &= \sum_{n=1}^{N} \int_{\boldsymbol{s} \in V_n} \sqrt{p(\boldsymbol{s} \mid \boldsymbol{s}_i) p(\boldsymbol{s} \mid \boldsymbol{s}_j)}\, d\boldsymbol{s} \\ &\approx \sum_{n=1}^{N} \sqrt{m_n(\boldsymbol{s}_i) m_n(\boldsymbol{s}_j)} \int_{\boldsymbol{s} \in V_n} p(\boldsymbol{s} \mid C_n)\, d\boldsymbol{s} \,. \end{aligned} \tag{A.2}$$

The second equation above is obtained by assuming that only the major cluster is contributing to its own Voronoi region so that the contributions from the others can be ignored. By the same reasoning, most of the probability mass of $p(\boldsymbol{s} \mid C_n)$ can be assumed to concentrate at $V_n$ so that $\int_{\boldsymbol{s} \in V_n} p(\boldsymbol{s} \mid C_n)\, d\boldsymbol{s}$ is nearly equal to one. Thus, the BC can be used to measure the closeness of the two color pixels $p(\boldsymbol{s} \mid \boldsymbol{s}_i)$ and $p(\boldsymbol{s} \mid \boldsymbol{s}_j)$ with $\boldsymbol{m}(\boldsymbol{s}_i)$ and $\boldsymbol{m}(\boldsymbol{s}_j)$ as

$$\psi(\boldsymbol{s}_i, \boldsymbol{s}_j) \approx \sum_{n=1}^{N} \sqrt{m_n(\boldsymbol{s}_i) m_n(\boldsymbol{s}_j)} = \sqrt{\boldsymbol{m}(\boldsymbol{s}_i)^T \boldsymbol{m}(\boldsymbol{s}_j)} \,. \tag{A.3}$$

Like conventional VQ, the accuracy of this approximation depends on the data distribution and number of clusters or Voronoi regions used. In our case, the Calinski-Harabasz index [6] is adopted as the criterion to determine the number of clusters, which was shown to yield good results in our computer experiments. More results on its performance can be found in the supplementary material.

## References


[1] J. Huang, S. R. Kumar, M. Mitra, W.J. Zhu, and R. Zabih, "Image indexing using color correlograms," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 1997, pp. 762–768.

[2] G. P. Qiu, "Indexing chromatic and achromatic patterns for content-based colour image retrieval," *Pattern Recognit*, vol. 35, no. 8, pp. 1675–1686, 2002.

[3] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE 7th Int. Conf. Comput. Vis.*, vol. 2. Ieee, 1999, pp. 1150–1157.

[4] E. Tola, V. Lepetit, and P. Fua, "Daisy: An efficient dense descriptor applied to wide-baseline stereo," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 5, pp. 815–830, 2010.

[5] S. C. Chan, C. W. Kok, K. T. Lai and K. L. Ho, "Codebook Generation and Search Algorithm for Vector Quantization using Arbitrary Hyperplanes," in *Proc. 1992 IEEE Workshop on Visual Signal Processing and Communications*, North Carolina, USA, 1992, pp. 74-79.

[6] T. Calinski and J. Harabasz, "A dendrite method for cluster analysis," *Communications in Stat.-theory and Methods*, vol. 3, no. 1, pp. 1–27, 1974.

[7] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *J. Royal Stat. Soc. Series C (Appl. Stat.)*, vol. 28, no. 1, pp. 100–108, 1979.

[8] P. C. Mahalanobis, "On the generalized distance in statistics," in *Proc. Nat. Inst. Sci.*, vol. 2, pp. 49–55, 1936.

[9] A. Bhattachayya, "On a measure of divergence between two statistical population defined by their population distributions," *Bulletin Calcutta Math. Soc.*, vol. 35, pp. 99–109, 1943.

[10] J. F. Wang, T. L. Zhang, and B. J. Fu, "A measure of spatial stratified heterogeneity," *Ecological Indicators*, vol. 67, pp. 250–256, 2016.

[11] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *Proc. 18th ACM Int. Conf. on Multimedia*, 2010, pp. 1469–1472.

[12] I. Kokkinos, M. Bronstein, and A. Yuille, "Dense scale invariant descriptors for images and surfaces," *Ph.D. dissertation*, INRIA, 2012.

[13] M. Maire, P. Arbelaez, C. Fowlkes, and J. Malik, "Using contours to detect and localize junctions in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. IEEE, 2008, pp. 1–8.

[14] M. Leordeanu, R. Sukthankar, and C. Sminchisescu, "Efficient closed-form solution to generalized boundary detection," in *European Conf. on Comput. Vis.* Springer, 2012, pp. 516–529.

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

[16] E. Trulls, I. Kokkinos, A. Sanfeliu, and F. Moreno-Noguer, "Dense segmentation-aware descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2013, pp. 2890–2897.

[17] T. Leung and J. Malik, "Contour continuity in region based image segmentation," in *European Conf. on Comput. Vis.* Springer, 1998, pp. 544–559.

[18] T. Brox and J. Malik, "Object segmentation by long term analysis of point trajectories," in *European Conf. on Comput. Vis.* Springer, 2010, pp. 282–295.

[19] C. Liu, J. Yuen, A. Torralba, J. Sivic, and W. T. Freeman, "Sift flow: Dense correspondence across different scenes," in *European Conf. on Comput. Vis.* Springer, 2008, pp. 28–42.

[20] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[21] Z. G. Li, X. M. Wu, and S. F. Chang, "Segmentation using superpixels: A bipartite graph partitioning approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012, pp. 789–796.

[22] X. F. Wang, Y. X. Tang, S. Masnou, and L. M. Chen, "A global/local affinity graph for image segmentation," *IEEE Trans. on Imag. Process.*, vol. 24, no. 4, pp. 1399–1411, 2015.

[23] C. Zhu, C. E. Bichot, and L. M. Chen, "Multi-scale color local binary patterns for visual object classes recognition," in *Proc. 20th Int. Conf. Pattern Recognit. (ICPR)*, 2010, pp. 3065–3068.

[24] X. F. Wang, H. B. Li, C. E. Bichot, S. Masnou, and L. M. Chen, "A graph-cut approach to image segmentation using an affinity graph based on ℓ 0-sparse representation of features," in *Proc. IEEE Int. Conf. on Imag. Process*, 2013, pp. 4019–4023.

[25] L. J. Sun and X. H. Liang, "Unsupervised image segmentation using global spatial constraint and multi-scale representation on multiple segmentation proposals," in *Proc. IEEE Int. Conf. on Imag. Process*, 2013, pp. 2704–2707.

[26] Microsoft Research Cambridge Object Recognition Image Database, 2015. [Online]. Available: https://www.microsoft.com/en-us/download/details.aspx?id=52644.

[27] T. Malisiewicz and A. A. Efros, "Improving spatial support for objects via multiple segmentations," in *British Mach. Vis. Conf. (BMVC)*, September 2007.

[28] 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," in *Proc. IEEE 8th Int. Conf. Comput. Vis.*, vol. 2, 2001, pp. 416–423.

[29] R. Unnikrishnan, C. Pantofaru, and M. Hebert, "Toward objective evaluation of image segmentation algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 929–944, 2007.

[30] M. Meil, "Comparing clusterings: an axiomatic view," in *Proc. ACM 22nd Int. Conf. Mach. Learn*, 2005, pp. 577–584.

[31] J. Freixenet, X. Munoz, D. Raba, J. Martí, and X. Cufí, "Yet another survey on image segmentation: Region and boundary information integration," in *European Conf. on Comput. Vis*, Springer, 2002, pp. 408–422.

[32] L. Zheng, S, Wang, Z, Liu, and Q, Tian. "Packing and padding: Coupled multi-index for accurate image retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 1939-1946.

[33] J. Van de Weijer, C, Schmid, J, Verbeek, and D, Larlus. "Learning color names for real-world applications," *IEEE Trans. on Imag. Process.*, vol. 18, no. 7, pp. 1512-1523, 2009.

[34] A. Mojsilovic. "A computational model for color naming and describing color composition of images," *IEEE Trans. on Imag. Process.*, vol. 14, no. 5 pp. 690-699, 2005.

The final version of record is available at http://dx.doi.org/10.1109/TIP.2018.2883580



[35] R. Benavente, M. Vanrell, and R. Baldrich, "Parametric fuzzy sets for automatic color naming," *J. Opt. Soc. Am. A. Opt. Image. Sci. Vis.*, vol. 25 no. 10 pp. 2582-2593, 2008.
[36] B. Berlin and P. Kay, *Basic color terms: Their universality and evolution*. Berkeley, CA: Univ. of California Press, 1969.
[37] P. Kay and C.K. McDaniel, "The linguistic significance of the meanings of basic color terms," *Language*, vol. 54, no. 3, pp. 610-646, 1978.
[38] R. Khan, J. van de Weijer, F. S. Khan, D. Muselet, C. Ducottet, and C. Barat. "Discriminative color descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2013, pp. 2866-2873.
[39] F. S. Khan, R. M. Anwer, J. van de Weijer, A. D. Bagdanov, A. M. Lopez, and M. Felsberg. "Coloring action recognition in still images," *Int. J. Comput. Vis.*, vol. 105, no. 3 pp. 205-221, 2013.
[40] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582-1596, 2010.
[41] M. E. Celebi, " Improving the performance of K-means for color quantization," *Image Vis. Comput.*, vol. 29, no. 4, pp. 260-271, Mar. 2011.
[42] G. Csurka, C. Dance, L. Fan, J. Willamowski and C. Bray, "Visual categorization with bags of keypoints," in *European Conf. on Comput. Vis.* Springer, 2004, pp. 1–2.
[43] F. Perronnin, and C. Dance. "Fisher kernels on visual vocabularies for image categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2007, pp. 1-8.
[44] H. Bay, T. Tuytelaars, and L. Van Gool, "Surf: Speeded up robust features," in *European Conf. on Comput. Vis.* Springer, 2006, pp. 404–417.
[45] B. Klein, G. Lev, G. Sadeh, and L. Wolf. "Fisher vectors derived from hybrid gaussian-laplacian mixture models for image annotation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015.
[46] F. Perronnin, J. Sanchez, and T. Mensink. "Improving the ´ fisher kernel for large-scale image classification," in *European Conf. on Comput. Vis.* Springer, 2004, pp. 143–156.
[47] V. Sydorov, M. Sakurada, and C. H. Lampert. "Deep fisher kernels–end to end learning of the fisher kernel gmm parameters," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014, pp. 1402-1409.
[48] J. Spanier and K. B. Oldham, *"The Basset* $\boldsymbol{K}_{\nu}(x)$ *."* Ch. 51 in An Atlas of Functions. Washington, DC: Hemisphere, pp. 499-507, 1987.
[49] E. Hellinger, "Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen," *Journal für die reine und angewandte Mathematik*, vol. 136, pp. 210-271. 1909
[50] A. Gersho, and R. M. Gray, *Vector quantization and signal compression*, Vol. 159. Springer Science & Business Media, 2012.
[51] L. Grady, "Space-variant computer vision: A graph-theoretic approach," *Ph.D. dissertation*, Dept. Cognit. Neural Syst., Boston Univ., Boston, MA, USA, 2004.
[52] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 171–184, Jan. 2013.
[53] E. Elhamifar and R. Vidal, "Sparse subspace clustering: Algorithm, theory, and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2765–2781, Nov. 2013.
[54] S. Kotz, T. Kozubowski and K. Podgorski, *The Laplace distribution and generalizations*, New York: Springer, 2001, p. 229-245.
[55] N. Kambhatla and T. Leen, "Dimension reduction by local principle component analysis," *Neural Comput.*, vol. 9, no. 7, pp. 1493-1500, Oct. 1997.
[56] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the Fisher vector: Theory and practice," *Int. J. Comput. Vis.*, vol. 105, pp. 222-245, 2013

**Zhouchi Lin** received the B.Eng. degree in automation from Sun Yat-sen University, Guangzhou, China, in 2012, the M.Eng. degree in Electrical and Computer Engineering from Cornell University, Ithaca, NY, USA, in 2013, and the Ph.D. degree from the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China, in 2018. His current research interests include multi-view image processing, image registration, and person re-identification.

**Hongdong Qin** received the B.Eng. degree in electronic and information engineering from The Hong Kong Polytechnic University, Hong Kong, China, in 2011. He is working towards Ph.D. in the Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China. His recent research focused on image and video coding.

**S. C. Chan (S'87 – M'92)** received the B.Sc. (Eng) and Ph.D. degrees from The University of Hong Kong, Pokfulam, in 1986 and 1992, respectively. Since 1994, he has been with the Department of Electrical and Electronic Engineering, the University of Hong Kong, where he is now a professor. His research interests include fast transform algorithms, filter design and realization, multirate and biomedical signal processing, communications and array signal processing, high-speed A/D converter architecture, bioinformatics, smart grid, and image-based rendering. He is currently a member of the Digital Signal Processing Technical Committee of the IEEE Circuits and Systems Society, and associate editors of Journal of Signal Processing Systems, Digital Signal Processing and IEEE Transactions on Circuits and Systems II. He was the chair of the IEEE Hong Kong Chapter of Signal Processing in 2000–2002, an organizing committee member of the 2003 IEEE ICASSP, the 2010 IEEE ICIP, and an associate editor of IEEE Transactions on Circuits and Systems I from 2008 to 2009.