January 2010

A method for recognizing the shape of a Gaussian mixture from a sparse sample set

H. J. Santos-Villalobos

M. Boutin

Santos-Villalobos, H. J. and Boutin, M., "A method for recognizing the shape of a Gaussian mixture from a sparse sample set" (2010). *Department of Electrical and Computer Engineering Faculty Publications.* Paper 47.
http://dx.doi.org/http://dx.doi.org/10.1117/12.848604

A method for recognizing the shape of a Gaussian mixture from a sparse sample set

Hector J. Santos-Villalobos and Mireille Boutin

School of Electrical and Computer Engineering, Purdue University
465 Northwestern Ave. West Lafayette, IN 47907 USA;

ABSTRACT

The motivating application for this research is the problem of recognizing a planar object consisting of points from a noisy observation of that object. Given is a planar Gaussian mixture model $\rho_T(\mathbf{x})$ representing an object along with a noise model for the observation process (the template). Also given are points representing the observation of the object (the query). We propose a method to determine if these points were drawn from a Gaussian mixture $\rho_Q(\mathbf{x})$ with the same shape as the template. The method consists in comparing samples from the *distribution of distances* of $\rho_T(\mathbf{x})$ and $\rho_Q(\mathbf{x})$, respectively. The distribution of distances is a faithful representation of the shape of generic Gaussian mixtures. Since it is invariant under rotations and translations of the Gaussian mixture, it provides a workaround to the problem of aligning objects before recognizing their shape without sacrificing accuracy. Experiments using synthetic data show a robust performance against type I errors, and few type II errors when the given template Gaussian mixtures are well distinguished.

Keywords: Bag of distances, Gaussian mixtures, Fingerprints, Information retrieval, Kolmogorov-Smirnov, Shape matching, Shape similarity

1. INTRODUCTION

The problem of searching for information is ubiquitous. In cyberspace, for example, many different engines allow users to find relevant webpages through a text query consisting of a few words. However, the information available in electronic format goes beyond text. Recognizing the relevant information when the query and/or the data being searched cannot be summarized as text is challenging. One of the main difficulties is the richness and complexity of the data, which often hinders our ability to focus on what is relevant for the search. Part of the problem is that there are often many different ways for a query to appear in the data but no effective way to remove this ambiguity from the data itself. Thus, a lot of research today is concerned with developing recognition method for non-text data (e.g. images, sounds, or videos), which match the speed and accuracy of text-based methods.

One problem of interest is shape recognition, that is to say the recognition of an object, such as a curve, surface, or volume, up to a rotation and translation of that object. As the ambiguity of representing a shape is well understood as a group transformation, shape recognition is a good starting point before attacking other recognition problems where the ambiguity cannot be parameterized explicitly.

Further author information: (Send correspondence to H.J.S.V.)
H.J.S.V: E-mail: hsantosv@purdue.edu, Telephone: 1 787 466 4460
M.B.: E-mail: mboutin@purdue.edu, Telephone: 1 765 494 3538

Computational Imaging VIII, edited by Charles A. Bouman, Ilya Pollak, Patrick J. Wolfe,
Proc. of SPIE-IS&T Electronic Imaging, SPIE Vol. 7533, 753305 · © 2010 SPIE-IS&T
CCC code: 0277-786X/10/$18 · doi: 10.1117/12.848604

The shape recognition paradigm we follow involves two key steps. 1) A representation for the shape of an object is obtained (e.g. a feature vector or even the object itself). 2) A comparison between the two representations is performed. In this paradigm, there is a tradeoff between complexity and faithfulness at the representation level, which translates into a tradeoff between speed and accuracy at the comparison level. However, some recent work[1] has shown that, in some cases, the tradeoff is nil for the vast majority of possible objects. For example, it was shown that the (unlabeled) pairwise distances of a point-set (called a *bag of distances*) is a faithful representation for all, but a set of measure zero of point-sets.

For translating this paradigm into practice, one must first understand how noise in the measurements affects the object, and subsequently its representation. One can then, either modify the comparison method, so it can deal with the noisy observed data, or find a way to estimate the shape representation from the data. For the case of an object represented by a point-set, a Gaussian mixture (GM) model can be used to represent the measurement noise. Previous work[2,3] has shown that the distribution $r(\Delta)$ of the Euclidean distance between two points drawn independently at random according to this GM generalizes the bag of distances concept while providing a faithful shape representation for the shape of generic GMs. In this paper, we propose a comparison method for the case where one GM is a known template and the other GM consists of an observed sparse set of points (one point per Gaussian). Such a setup occurs, for example, in the problem of fingerprint identification using minutiae, where an affine analysis can provide the parameters of the GM, and the fingerprint query is performed inline using observed minutiae. The comparison method we proposed is fast enough to be executed in real time, and as our experiments indicate, its accuracy is very good.

The rest of this paper is divided into three sections. The following section summarizes the theoretical results on which our approach is based. Section 3 contains our proposed comparison method. Then, Section 4 presents a numerical evaluation of the method's accuracy. We conclude in Section 5.

2. THEORETICAL BACKGROUND

The fact that distributions of invariants can be used to faithfully represent objects modulo some group actions was proved by Boutin and Kemper,[1] for the case of point-sets. In particular, the set of pairwise distances of a point-set, what we call the Bag of Distances (BoD), was shown to be a lossless representation for the vast majority of point-sets. More specifically, the following theorem was proved. See [2] for a simple proof.

Theorem 1. There exists a polynomial f in $2n$ variables such that if the points $p_1, p_2, \ldots, p_n \in \mathbb{R}^2$ satisfy $f(p_1, p_2, \ldots, p_n) \neq 0$, then for any other point-set $\bar{p}_1, \bar{p}_2, \ldots, \bar{p}_n$ having the same bag of distances as that of $p_1, p_2, \ldots, p_n$, there exists an orthogonal matrix $M \in \mathbb{R}^{2\times 2}$, a translation vector $T \in \mathbb{R}^2$ and a permutation $\pi \in S_n$ such that

$$\bar{p}_i = M p_{\pi(i)} + T, \text{ for all } i = 1, \ldots, n.$$

The point-sets that do not lie on the zero-set of the aforementioned polynomial f are called *generic* point-sets.[2]

Some recent work aims to generalize Thm. 1 to the case of non-deterministic point-sets. In particular, the problem of representing the shape of a GM $\rho(\mathbf{x})$ was considered. In 3, the probability density function $r(\Delta)$ of the distance Δ between two points x_1 and x_2 drawn independently from $\rho(\mathbf{x})$ was proposed as a representation. In 2, $r(\Delta)$ was shown to be a faithful representation of the shape of the vast majority of planar GM. More specifically, the following theorem was proved.

Theorem 2.[3] Suppose that two Gaussian Mixtures $\rho(\mathbf{x})$, $\bar{\rho}(\mathbf{x})$ are such that their respective means forms a generic point-set. Then $\rho(\mathbf{x})$ and $\bar{\rho}(\mathbf{x})$ have the same distribution of distances, $r(\Delta) = \bar{r}(\Delta)$, if and only if they have the same shape, i.e. if and only if there exists an orthogonal matrix $M \in \mathbb{R}^{2\times 2}$ and a translation vector $T \in \mathbb{R}^2$ such that

$$\rho(\mathbf{x}) = \bar{\rho}(M\mathbf{x} + T).$$

Theorem 2 states that if $r(\Delta) = \bar{r}(\Delta)$, then $\rho(\mathbf{x}) \equiv \bar{\rho}(\mathbf{x})$. Therefore, we can avoid the difficult task of aligning the GMs by comparing the distribution of distances $r(\Delta)$ and $\bar{r}(\Delta)$.

3. THE PROPOSED METHOD

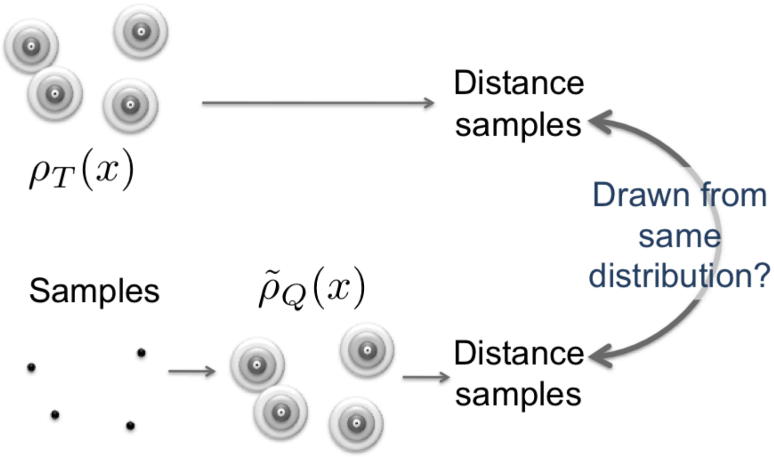


Figure 1. Flowchart of the tasks performed by the proposed method.

Given is a sparse set of points $p_1, p_2, \ldots, p_n \in \mathbb{R}^2$ drawn from a Gaussian mixture $\rho_Q(\mathbf{x})$ with n Gaussian components, each with standard deviation matrix $\sigma^2 \cdot \mathbb{I}_{2\times 2}$. The parameters n and σ are known, but the remaining parameters of $\rho_Q(\mathbf{x})$ are unknown. We assume that the samples were drawn from distinct Gaussian components of $\rho_Q(\mathbf{x})$. Also, given is a template Gaussian mixture $\rho_T(\mathbf{x})$ with n Gaussian components, each with standard deviation matrix $\sigma^2 \cdot \mathbb{I}_{2\times 2}$.

Our goal is to determine if $\rho_T(\mathbf{x})$ and $\rho_Q(\mathbf{x})$ have the same shape, that is to say if $\rho_T(\mathbf{x}) = \rho_Q(M\mathbf{x} + T)$, for some rotation matrix $M \in \mathbb{R}^{2\times 2}$ and translation vector $T \in \mathbb{R}^2$. The method we propose contains five steps, namely.

Step 1: Use $p_1, p_2, \ldots, p_n$ to obtain an approximation $\tilde{\rho}_Q(\mathbf{x})$ of $\rho_Q(\mathbf{x})$.

Step 2: Draw N independent samples $d_1, d_2, \ldots, d_N$ from $r(\Delta)$, the distribution of distances of $\rho_T(\mathbf{x})$, and draw N independent samples $\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_N$ from $\tilde{r}(\Delta)$, the distribution of distances of $\tilde{\rho}_T(\mathbf{x})$.

Step 3: Measure the likelihood if the samples $d_1, d_2, \ldots, d_N$ and the samples $\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_N$ were drawn from the same distribution.

Step 4: Repeat step 2 and step 3 a total of K times.

Step 5: Final decision.

Below we describe each step in details.

Step 1: Approximation of $\rho_Q(\mathbf{x})$.

Since the set of sample points $p_1, p_2, \ldots, p_n$ is sparse, and the points were drawn from distinct Gaussian mixture components, we set

$$\tilde{\mu}_i = p_i, \quad i = 1, 2, \ldots, n$$

Then, we set $\tilde{\sigma} = \lambda\sigma$. As we set the standard deviation of the components $\Sigma_i = \lambda^2\sigma^2 \cdot \mathbb{I}_{2\times 2}$, all the parameters of $\tilde{\rho}_Q(\mathbf{x})$ are then defined. In our numerical experiments, (See Section 4) we found that $\lambda \in [0.66, \ldots, 1.33]$.

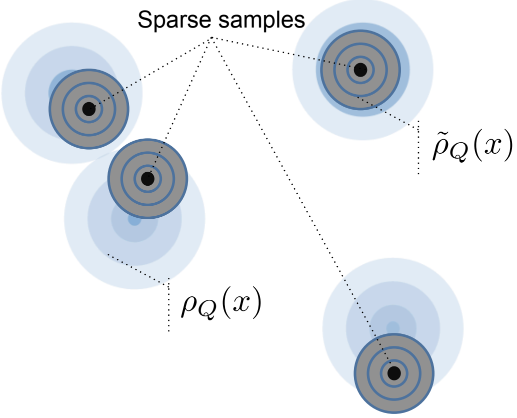


Figure 2. Approximation of $\rho_Q(\mathbf{x})$ from the given point samples.

As roughly 95% of the samples drawn from a Gaussian distribution will fall within a distance 2σ of its mean, the regions of high density of each of the components of $\tilde{\rho}_Q(\mathbf{x})$ are likely to overlap with those of $\rho_Q(x)$, and these $\tilde{\rho}_Q(\mathbf{x})$ and $\rho_Q(\mathbf{x})$ are likely to be similar (See Fig. 2)

Step 2: Sampling of the two distributions of distances.

To obtain independent samples from a distribution of distances of a Gaussian mixture (either $\tilde{\rho}_Q(\mathbf{x})$ or $\rho_T(\mathbf{x})$), we draw two independent samples $\mathbf{x}_1$ and $\mathbf{x}_2$ from the given Gaussian mixture and measure their Euclidean distance $|\mathbf{x}_1 - \mathbf{x}_2|_{L_2}$. We do this a total of N times in order to obtain N independent distance samples. The distance samples obtained from $\rho_T(\mathbf{x})$ are labeled $d_1, d_2, \ldots, d_N$ and those from $\tilde{\rho}_Q(\mathbf{x})$ are labeled $\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_N$.

Step 3: Evaluation of the p-value of the hypothesis.

In order to decide if the distance samples $d_1, d_2, \ldots, d_N$ and the distance samples $\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_N$ were drawn from the same distribution, we use the Kolmogorov-Smirnov (KS) test. The KS test[4,5] is a statistical test that quantifies the dissimilarities between two sample sets. More precisely, the KS test measures the distance

$$D^* = \max_{\Delta \in \mathbb{R}} \left(|R_T(\Delta) - R_Q(\Delta)| \right),$$

where $R_T(\Delta)$ and $R_Q(\Delta)$ are the cumulative distributions of the sample sets $\{d_1, d_2, \ldots, d_N\}$ and $\{\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_N\}$, respectively. The quantity D^* is then used to estimate the likelihood (p-value) that the null hypothesis is true at a 5% significance level; the null hypothesis being that the distance samples were drawn from the same distribution. Type I errors (i.e. rejecting the null hypothesis when it is true) occur when the method incorrectly finds that the shape of the query $\rho_Q(\mathbf{x})$ matches that of the template $\rho_T(\mathbf{x})$. Conversely, type II errors (i.e. accepting the null hypothesis when it is false) occur when the method incorrectly finds that the shape of the query and the template differ.

The K-S test was preferred after considering the following advantages: 1) It is not required to bin the samples like in the χ^2 test. 2) The test compares the samples directly. 3) We don't need a prior knowledge of the distribution of the samples or any statistical parameters. 4) The K-S test is robust against small sampling sizes.[6] However, the test tends to be more sensitive near the center of the distribution than the tails.[6] Therefore, we confirmed the simulations results with the Kuiper statistical test.[7] The Kuiper test is a modified K-S test, which gives equal importance to all regions of the distribution.[6, 7] The results with the Kuiper test confirmed that the inferior sensitivity of the K-S test on the tails of the distribution does not affect the effectivity of the proposed method.

Step 4: Repeat step 2 and step 3 a total of K times.

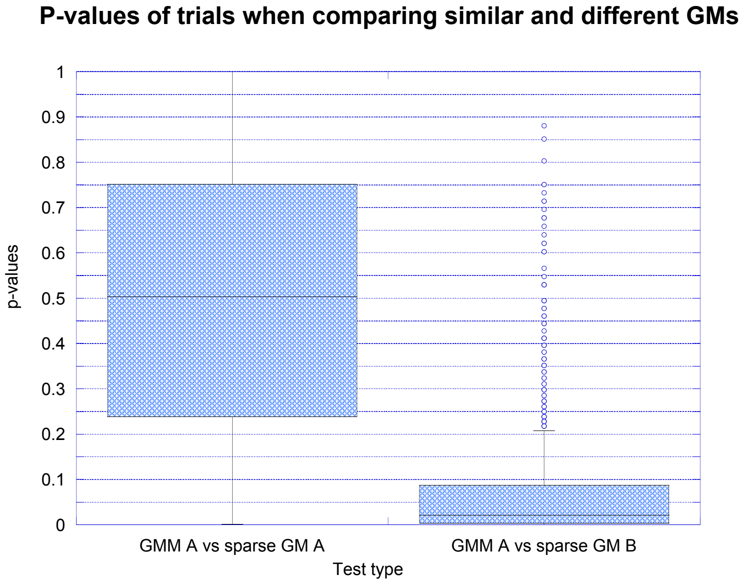
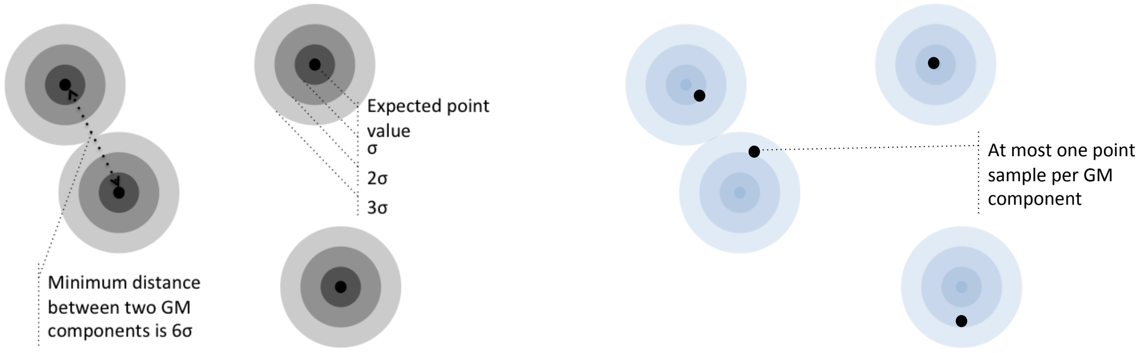


Figure 3. Box-and-whisker diagrams for p-values when comparing two BoDs that belong to the same distribution $r(\Delta)$ and two BoDs that belong to different distributions.

As the distance samples are obtained independently at random, it is possible that they are "bad" samples, in the sense that they do not approximate the distribution of distances well. To mitigate this, we generate K different sample sets and obtain the p-value for each set. Figure 3 illustrates the variability of the p-value for different trials.

Step 5: Final decision.

Our final decision is based on the median p-value of all the K trials performed. As most distance sample sets are relatively "good" (See Fig. 3), a few hundred trials are sufficient.



(a) (b)

Figure 4. Illustrations of Gaussian mixtures used to test our method. (a) The template Gaussian mixtures are chosen so that their components do not overlap too much. (b) The point samples $p_1, p_2, \ldots, p_n$ are drawn from distinct Gaussian mixtures components.

4. NUMERICAL EXPERIMENTS

Before introducing the experimental results, we describe the Gaussian mixtures used to assess the method. Each template Gaussian mixture $\rho_T(\mathbf{x})$ is generated with the following criteria: 1) The mixture is constrained to a 500×500 square grid with 64-bits of floating-point precision. 2) All the components in the mixture share the same σ. In our simulations, σ was fixed to three. 3) To avoid overlapping between the Gaussian mixture components, the distance between the centers of any two components is constrained to a minimum of six times their standard deviation σ (See Fig. 4a). Each component's location (i.e. the mean μ_i) is generated with a uniform distribution pseudo-random number generator. A new random location is generated until the new component meets the previous criteria. The number of GM components n was taken to be $n = 25, 40, 55, 70, 85, 100$. Finally, ten $\rho_T(\mathbf{x})$ were generated for each particular configuration, for a total test set of 60 synthetic Gaussian mixtures.

Sparse point sets are generated from the Gaussian mixture templates in the synthetic set. As shown in Fig. 4b, to form a sparse set, a single point sample is drawn from each component in the source template. Noise is added to the sparse point samples by changing the template's standard deviation σ by a noise value σ_{noise}, where $\sigma_{noise} = \gamma\sigma$ and $\gamma \in [0, \ldots, 5.33]$. The purpose of σ_{noise} is to constraint the distance between each point sample p_i and the center μ_i of the corresponding mixture component. Consequently, a sparse set generated without noise (i.e. $\sigma_{noise} = 0$) has its point samples $p_1, \ldots, p_n$ located at $\mu_1, \ldots, \mu_n$, respectively. As σ_{noise} increases, the point samples in the sparse set may be located further away from the components' centers (i.e. deforming the shape).

The simulation procedure is described by the following 5 steps. 1) A sparse set of samples is generated from a modified Gaussian mixture $\rho_Q(\mathbf{x})$ with noise σ_{noise}. 2) A template Gaussian mixture $\rho_T(\mathbf{x})$ with standard deviation σ is obtained from the GM synthetic set. 3) The query Gaussian mixture $\tilde{\rho}_Q(\mathbf{x})$ is generated with standard deviation $\tilde{\sigma}$, where $\tilde{\sigma} = \lambda\sigma$. 4) Two sets of distance samples $d_1, d_2, \ldots, d_N$ and $\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_N$ are drawn from $\rho_T(\mathbf{x})$ and $\tilde{\rho}_Q(\mathbf{x})$, respectively, and compared with the proposed method. 5) If the templates $\rho_T(\mathbf{x})$ and $\rho_Q(\mathbf{x})$ are equivalent, the method should return a p-value ≥ 0.05; otherwise a type I error occurs. Conversely, if the templates are distinct, the method should return a p-value < 0.05; otherwise a type II error occurs.

Figure 5 shows the error rates of the proposed method. The vertical axis show the error rate scale, while the

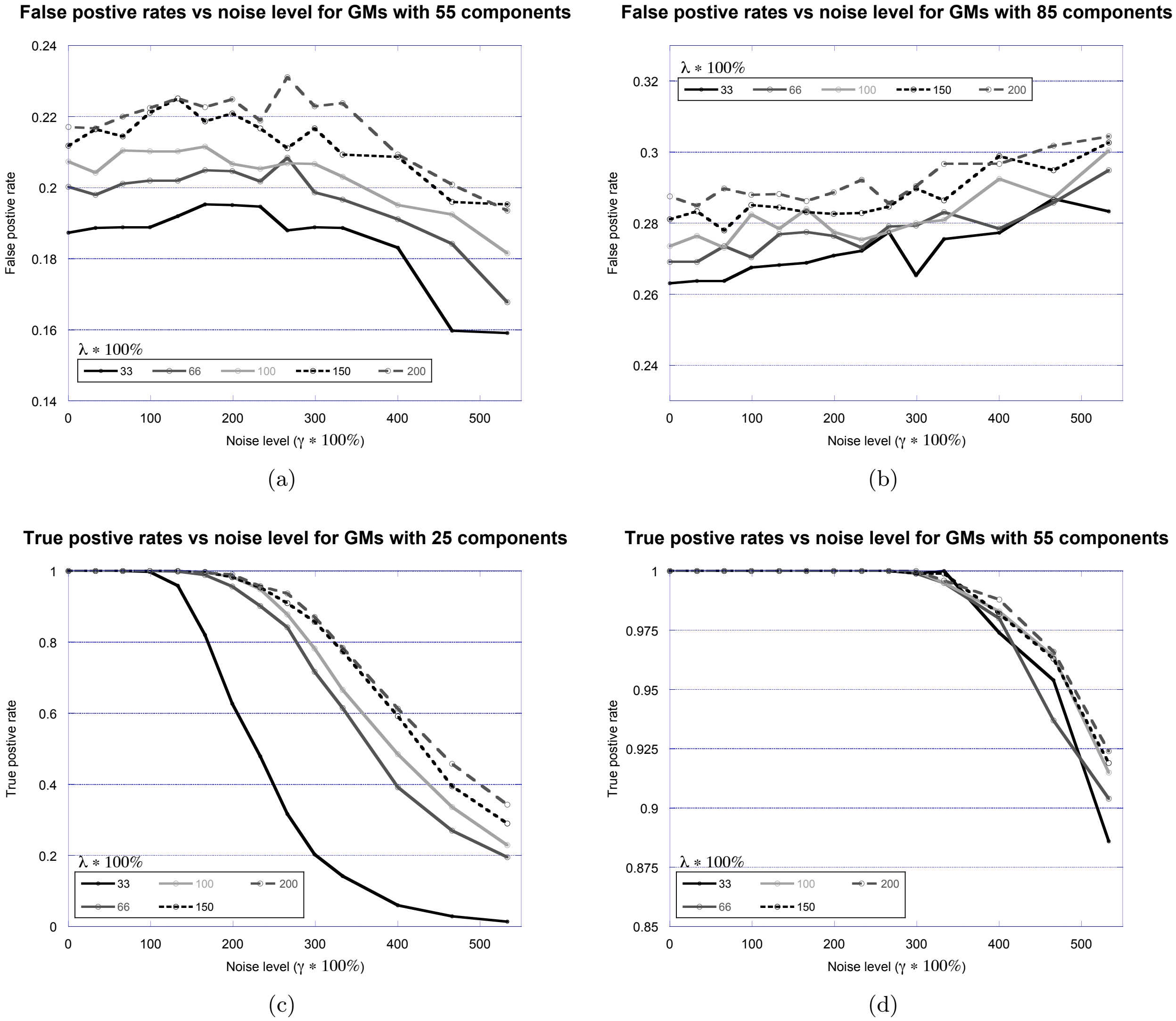


Figure 5. Error rate of our method when comparing synthetic Gaussian mixtures and sparse sample sets. (a) False positive rates for mixtures with 55 components. (b) False positive rates for mixtures with 85 components. (c) True positive rates for mixtures with 25 components. (d) True positive rates for mixtures with 55 components.

horizontal axis shows the noise level γ. In addition, each line shows the results for different values for λ.

Overall the proposed comparison method accurately classified the synthetic shapes–type I and II errors were minimal. However, as shown in Figure 5, the method produced a significant amount of classification errors for a particular set of Gaussian mixture configurations. Specifically, Figures 5a and 5b show the error rates when comparing Gaussian mixtures with 55 and 85 components respectively. Mixtures with a large number of components congest the 500 × 500 plane. Therefore, it is difficult to generate a complete set of templates with noticeable differences. In contrast, mixtures in the synthetic set with a fewer components have evident dissimilar shapes. As expected, the proposed method did not produce type II errors for that subset of mixtures.

There are two parameters that have an effect on the method's robustness against noise. As shown in Figures 5c and 5d a larger λ reduced type I errors, while Figures 5a and 5b show that it also increases the number of type II errors. Another interesting result is that the robustness against noise increases as the number of components in the mixture also increases. Observe in Figures 5d and 5c that increasing the number of components drastically reduced the number of type I errors, while the type II errors in Fig. 5b almost remain constant for the same σ. When the dissimilarity between two shapes is evident, a larger number of components implies a more difficult shape pattern to match. Consequently, even when noise is added, it is difficult to find a match between different shapes. From our analysis of the experimental results, we concluded that appropriate values for λ are between 0.66 and 1.3.

5. CONCLUSION AND FUTURE WORK

We presented a method to determine whether a set of points $p_1, p_2, \ldots, p_n \in \mathbb{R}^2$ is a noisy observation (up to a rotation and translation) of a planar object consisting of a point-set. For generic point-sets whose observation noise can be modeled as a Gaussian mixture, the method reduces the dimensionality of the problem from two to one without compromising accuracy. This is done by considering the distribution of distances of the underlying Gaussian mixture, a shape representation that has been shown to be faithful for most Gaussian mixtures.[2,3] Since distances are unchanged by a rotation/translation, our method removes the problem of object alignment.

The comparison method comprises five steps: 1) Use $p_1, p_2, \ldots, p_n$ to obtain an approximation $\tilde{\rho}_Q(\mathbf{x})$ of $\rho_Q(\mathbf{x})$. 2) Draw N independent samples $d_1, d_2, \ldots, d_N$ from $r(\Delta)$, the distribution of distances of $\rho_T(\mathbf{x})$, and draw N independent samples $\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_N$ from $\tilde{r}(\Delta)$, the distribution of distances of $\tilde{\rho}_T(\mathbf{x})$. 3) With the KS test, measure the likelihood that the samples $d_1, d_2, \ldots, d_N$ and the samples $\tilde{d}_1, \tilde{d}_2, \ldots, \tilde{d}_N$ were drawn from the same distribution. 4) Repeat Step 2 and Step 3 a total of K times. 5) Use the median p-value of the K trials to make the final decision. If the median is above 0.05, we conclude that the objects have the same shape, i.e. $\rho_T(\mathbf{x}) = \rho_Q(M\mathbf{x} + T)$ given a rotation matrix M and a translation vector T. Otherwise, we conclude that the objects have a different shape.

Our empirical assessment of the method with synthetic data showed the potential of the method. The comparisons did not produced Type I errors unless high noise levels were added to the sparse mixtures. Also, Type II errors were only significant when the number of GM components was large, so the objects congested the 500×500 grid. This is expected, as it is difficult to generate dissimilar shapes when the GM components occupy most of the finite plane.

In future work, it would be interesting to study GMs whose components overlap and with varying standard deviation.

ACKNOWLEDGMENTS

This work was supported in parts by NSF grant CCF-0728929.

REFERENCES

[1] Boutin, M. and Kemper, G., "On reconstructing n-point configurations from the distribution of distances or areas," *Adv. Appl. Math.* **32**, 709–735 (2004).

[2] Lee, K., Boutin, M., and Comer, M., "Lossless shape representation using invariant statistics: the case of point-sets," in [*Proc. Asilomar Conf. Sig., Syst., and Comp.*], (October 29-November 1 2006).

[3] Boutin, M. and Comer, M., "Faithful shape representation for 2d gaussian mixtures," in [*IEEE Int'l Conference on Image Processing (ICIP)*], (September 16-19 2007).

[4] Massey, Frank J., J., "The kolmogorov-smirnov test for goodness of fit," *Journal of the American Statistical Association* **46**(253), 68–78 (1951).

[5] Stephens, M. A., "Use of the kolmogorov-smirnov, cramer-von mises and related statistics without extensive tables," *Journal of the Royal Statistical Society* **32**(1), 115–122 (1970).

[6] Press, W. H., Teukolsky, S. A., Vetterling, W. T., and Flannery, B. P., [*Numerical Recipes in C: The art of Scientific Computing*], Cambridge University Press (1992).

[7] Kuiper, N. H., "Tests concerning random points on a circle," in [*Proceedings of the Koninklijke Nederlandse Akademie van Wetenschappen*], 38–47 (1962).