J. Indones. Math. Soc.
Vol. 16, No. 1 (2010), pp. 51–67.

THE NEW NOTION DISTANCE OF CONTENT BASED IMAGE RETRIEVAL (CBIR)

Dyah E. Herwindiati[1], Sani M. Isa[2], and Rahmat Sagara[3]

[1]**Department of Informatics Engineering - Universitas Tarumanegara, dyah.fti.untar@gmail.com**
[2]**Department of Informatics Engineering - Universitas Tarumanegara, sani.fti.untar@gmail.com**
[3]**Department of Mathematics Education - Sampoerna School of Education, rahmat.sagara@sampoernaeducation.ac.id**

Abstract. This paper proposes a new notion distance on the CBIR process which is derived from the measure of multivariate dispersion called vector variance (VV). The minimum vector variance (MVV) estimator is robust estimator having the high breakdown point. The CBIR is a retrieval technique using the visual information by retrieving collections of digital images. The process of retrieval is carried out by measuring the similarity between query image and the image in the database through similarity measure. Distance is a metric often used as similarity measure on CBIR. The query image is relevant to an image in the database if the value of similarity measure is 'small'. This means that a good CBIR retrieval system must be supported by an accurate similarity measure. The classical distance is generated from the arithmetic mean which is vulnerable to the masking effect. The appearance of extreme data causes the inflation of deviation of the arithmetic mean, this implies the distance between the extreme data or the outlier becomes closer than it supposed to be. In the end of section we discuss the high performance of the MVV robust distance to CBIR process.

Key words and Phrases: High Breakdown Point, Image Retrieval, Query image, Similarity Measure, Image Visual Content

2000 Mathematics Subject Classification: 62H35
Received: 11-01-2010, accepted: 25-03-2010.

Abstrak. Karya ilmiah ini mengusulkan suatu ukuran jarak yang baru pada proses CBIR yang diturunkan dari ukuran dispersi multivariate yang dikenal dengan vektor varians (VV). Minimum vektor varians (MVV) adalah robust estimator yang mempunyai *breakdown point* tinggi. Metode CBIR adalah teknik untuk melakukan pencarian citra digital berdasarkan informasi visual dalam membangun sistem pencarian untuk data berkapasitas besar. Sistem pencarian CBIR yang baik harus didukung oleh nilai *similarity measure* yang akurat. *Similarity measure* adalah suatu teknik yang digunakan untuk mengukur tingkat kesamaan antara citra kueri dengan citra pada basis data berdasarkan jarak antara dua citra tersebut. Suatu citra kueri disebut relevan terhadap citra pada basis data jika nilai *similarity measure* dari citra tersebut 'kecil'. Beberapa ukuran jarak yang sering digunakan sebagai ukuran kesamaan proses CBIR adalah jarak klasik yang dibangun dari *aritmetic mean* dan rentan terhadap *masking effect*. Sebuah data ekstrim dapat menyebabkan membesarnya deviasi observasi terhadap rata-ratanya, sehingga jarak antara nilai pusat terhadap rata-rata tersebut menjadi lebih dekat. Di akhir pembahasan karya ilmiah ini ditunjukkan kinerja yang baik dari jarak robust MVV untuk proses CBIR.

Kata kunci: *Breakdown Point* Tinggi, *Image Retrieval*, Citra Kueri, *Similarity Measure*, *Image Visual Content*

1. Introduction

Informations contained in an image can be visual informations or semantic informations. The visual informations can be stated in general contexts in form of colours, textures, shapes, spatial relations, or in other specified forms which valid in the domain of certain problems. Content-based image retrieval (CBIR) is a retrieval technique which uses the visual information by retrieving collections of digital images. The visual information are then extracted and stated as a feature vector which in the sequel then forms a feature database.

Content-based image retrieval (CBIR), is also known as query by image content (QBIC) and content-based visual information retrieval (CBVIR) is an application of the computer vision to the problem of digital pictures retrieval in a large data base. This retrieval system is proposed as an attempt in providing better pictures management and retrieval, which traditionally uses explanation texts of the pictures. The process of retrieval is carried out by measuring the similarity between query image and the image in the database through similarity measure.

Similarity measure is a method to measure the distance between the feature vector of query image and the feature vector of image in the database. The value of similarity measure shows the level of similarity between the images. An image is called 'similar' with an image in the database if the value of similarity measure is 'small'. This means that a good retrieval process must be supported by an accurate similarity measure.

Distance is a metric which is often used as similarity measure on CBIR. The classical distance is often used in the retrieval process. The distance is generated from an arithmetic mean. The precission of classical distance is not high because of the masking effect. The appearance of extreme data causes a large enough deviation of the arithmetic mean, this implies that the masking effect causes the extreme data or the outlier becomes closer to the clean data. Barnet and Lewis [4] stated that the masking effect is caused by the tendency of extreme data which are not recognized as outliers, which in fact they are outliers. The data even covers other extreme data so they cannot be detected as outliers.

Various robust estimation methods have been proposed to reduce the masking effect. Recently, the most popular approach, is a robust approach that uses a criteria which minimize the determinant of covariance matrix, this is well known as minimum covariance determinant (MCD). Some researchers, for example Rousseeuw and Van Zomeren [21], Hadi [9], Hawkins [12], Becker & Gather [5], Rousseeuw and Van Driessen [22], Billor et.al [6] gave interesting and important information about sub robust sample of MCD. The sub sample is used to compute robust estimator which is useful to generate a robust distance.

MCD is derived from the multivariate dispersion measure which is known as generalized variance (GV). The word dispersion synonyms to the word spread which is shown by a group of data (Anderson [2]). The most popular measure of dispersion in univariate data is the standard of deviation. Different from univariate data, the measure of dispersion on the multivariate data is generated based on the structure of covariance which is recorded in a matrix of covariance $\mathbf{\Sigma}$. There are two popular multivariate dispersion, the total variance (TV), and the generalized variance (GV).

In the development of the multivariate analysis, GV is more popular than TV. This is very natural, since GV or $|\mathbf{\Sigma}|$ involves all of the information in every element of $\mathbf{\Sigma}$, no matter those is contained in the variance structure or covariance. Meanwhile the total variance (TV), which is defined as $\mathrm{Tr}(\mathbf{\Sigma})$ does not involve the covariance structure. TV only involves the variance structure.

This paper proposes robust distance on the CBIR process which is derived from the measure of multivariate dispersion called vector variance (VV). Good properties of VV mentioned by Herwindiati et.al [13] that VV takes shorter time of computation than CD. Further description of VV will be given on Section 3. Two examples in Section 4 discuss the CBIR process on the image of a flower based on feature of color and feature of texture. Section 4 also discusses the performance of MVV through the precision of retrieval process. We hope that the discussion on this paper will contribute to the development of CBIR.

2. Image Retrieval

Content based image retrieval (CBIR) is a technique uses the visual content to retrieve images from a large scale, in line with characteristics desired by users. CBIR has been an interesting topic of research that attracts many researchers since

the early of 90's. In the last decade, a lot of progress attained in theoretical research CBIR or in the development of the CBIR system. But, up to now there are still many challenging problems in the field of CBIR which attracts attention of many scientists from various disciplines.

Late of 1970's is the beginning era of research in CBIR. On 1979 there was conference on application of database technique in image, held in Florence. Since then, the application of image database management technique became an area of research that attracted many scientists. The technique used in the beginning of CBIR did not used the visual content yet, but relied on the textual information of each image (Text Based Image Retrieval – TBIR). On the other word, additional textual information is needed on every image before retrieval. Then the images can be retrieved by using the textual approach DBMS. Text based image retrieval uses the traditional database technique to manage the image database. Through textual description, image can be organized based on topic or level of hierarchies to make the navigation process and browsing easier. One thing which is almost impossible is to create a system that able to give additional textual information to every image, so most TBIR system needs manual process to add the textual information. This additional process is very difficult to carry on when the size of image database is too large, and it is very subjective, sensitive to the context and incomplete.

In 1992, National Science Foundation of United States held a workshop on Information Management System which aimed to determine future direction of the system of image database management. At that time researchers had already realized a fact that taking advantage of the visual properties contained in images is a more efficient way and intuitive in representing and indexing visual information of images. Scientist from various backgrounds such as computer vision, database management, human-computer interface, and information retrieval since then started their research in CBIR. Since 1997, publication in visual information extraction, organization, indexing, user query and interaction, and database management has increased significantly. An advance in CBIR was marked by commercial development of image retrieval system for government organization, private institutions and hospitals.

CBIR makes use of visual contents of image, such as shapes, texture, spatial layout, to represent and to index images. Generally in an image retrieval system, the visual content of images is stored in a multi dimension feature vector. To retrieve an image, users enter inputs of the form image query or sketch. Then the CBIR system computes the feature vector of the query images or sketch. Similarity between the feature vectors of query image/sketch is obtained based on a measure of distance or index scheme. The index scheme is a more efficient way to retrieve images on the database. The latest CBR system has entered feedbacks from users (relevance feedback) to alter the retrieval process to obtain better result, perceptually and to get more precise semantic meaning. Long [17] describes the CBIR process as Figure 1.

A good visual content descriptor must be invariant to the variance caused by the process of image formation or local. The global descriptor makes use visual

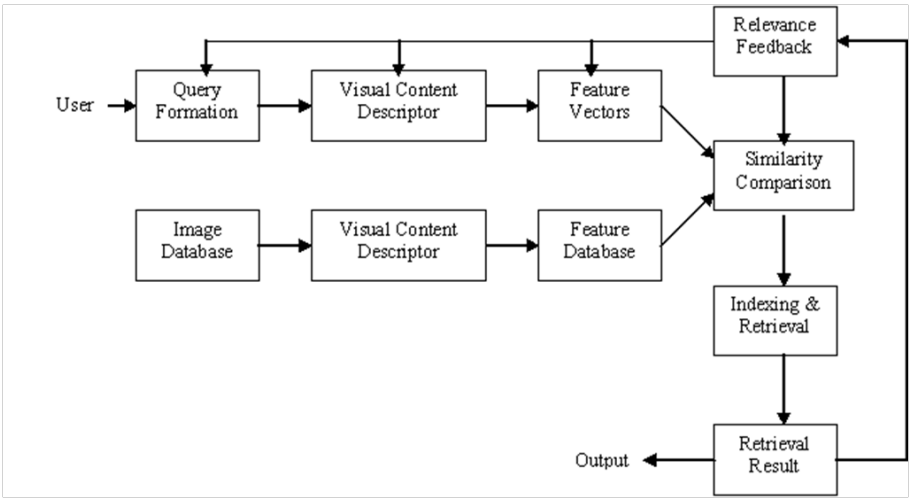

FIGURE 1. The Diagram of CBIR

information from the whole images, meanwhile the local descriptor makes use visual information of image region to describe the visual content of images. A system of image retrieval generally stores visual contents of an image in a multi dimension feature vector. In an image retrieval process, users enter inputs of the form query image. Next, the CBR system computes the feature vector of the query image. Similarity between the feature vectors of the query images is obtained based on the measurement of distance of index scheme. This research uses the distance measure as a measurement tool of similarity of global visual content.

3. The Robust Distance for Similarity OF CBIR

Basically almost every data set contains anomalies data or labeled outlier of various percentages. Outlier occurs in a data set because of unusual events, such as experiment failure, instrument damage, and so on. Outlier influences the data analysis that uses statistical techniques. The appearance of one or more outliers will lead to distortion of sample mean and variance so it causes faulty conclusions.

Barnet and Lewis [4] define an outlier to be one or more inconsistent observations. The word 'inconsistent' on the definition is not easy to be formulated in general situations. This reason makes people, up to now; develop better methods in identifying outliers. For instance, in the univariate case, Irwin [16] proposed that the deviation of the mean as the criteria of outlier, Thomson [23] developed Irwin's idea [16] by proposing a new measuring tool, i.e. the ratio between the deviation from its mean and sample's standard deviation. The Statistics proposed by Thomson [23] apparently has a very big impact to further development. Today, outlier detection can be found in every work especially when statistical techniques

are applied to the multivariate data cases. Even for data mining of large data and high dimension, such as in data mining and knowledge discovery (Angiulli and Pizzuti, [3]), and intrusion detection (Ye et al., [27]), the mechanism of detection outlier is extremely important parts of a thorough analysis.

Derquenne [7] implicitly stated the measure of identification multivariate outliers is created by a technique transforms random vectors to be random variables so that candidates of outlier will be seen more clearly. The most popular transformation is the Mahalanobis distance.

Mahalanobis distance is a distance which measures every observation $\vec{X}$ to $\vec{\bar{X}}$ given by sample covariance $\boldsymbol{S}$, and formulated by

$$d^2_{\boldsymbol{S}}(\vec{\bar{X}}, \boldsymbol{S}) = (\vec{X} - \vec{\bar{X}})^t \boldsymbol{S}^{-1} (\vec{X} - \vec{\bar{X}}).$$

Presence of one or more outliers alters the arithmetic mean $\vec{\bar{X}}$ significantly, and $d^2_{\boldsymbol{S}}(\vec{\bar{X}}, \boldsymbol{S})$ increases. The Mahalanobis distance is not robust to outliers and vulnerable to masking effect. Barnet and Lewis [4] define the masking effect as an effect of the tendency of extreme data which is not recognised as outliers, but in fact are outliers.

One of impacts of masking effect is occurrence of distortion of sample mean and variance. To handle this problem, the method of robust estimator introduced by Huber [14] is applicable as theoretical foundations of the construction of distance which is robust Mahalanobis.

Various methods of robust estimation can be found in literatures. Rousseeuw [19] introduced minimum volume ellipsoid (MVE) and minimum covariance determinant (MCD) methods with $h = [\frac{n}{2} + 1]$ and n is the size of the sample. Here $[z]$ represents the integer part of z. Rousseeuw and van Zomeren [21] proposed the use of MVE to choose subsets having minimum volume of ellipsoid and covers at least h data. Hawkins [12] introduced the feasible solution algorithm (FSA) to determine h data which give covariance matrices of minimum determinant. For the same sake, Rousseeuw and van Driessen [22] proposed FMCD. The difference between the two methods lies in the process of determining h data. It was mentioned by Hardin and Rocke [10], also Werner [24] that FMCD is faster than MVE, MCD or even than FSA.

FMCD has very impressive algorithm efficiency (Werner, [24]). But, according to the author, this thing happens only on multivariate data of low dimension. For large data of higher dimension, the efficiency of the FMCD algorithm is worsening. This is due to computations of the determinant of covariance matrices, which takes time of order $O(p^3)$ by Cholesky's method. Here p is the number of variables.

As a measure of dispersion, CD has wider applications than TV, though there is limitedness of CD. The main limitedness is its property, CD = 0 if there is variable of zero variance or if there is a variable which is a linear combination of others. Therefore the author proposes another measure to detect multivariate outlier, this called minimum vector variance (MVV).

MVV is generated by measure of disperse multivariate vector variance (VV). Different than CD, VV is still able to measure the multivariate dispersion, though, the covariance matrix is singular. Geometrically VV is a square of length of the diagonal of a parallelotope generated by all principal components of $\vec{X}$, Djauhari [8]. This good property of VV motivated Herwindiati et al [13] to generate a new measure, MVV, as a measure to detect outliers. Relation among TV, GV or CD, and VV can be described as follow: Suppose $\vec{X}$ is a random vector of covariance matrix $\boldsymbol{\Sigma}$ of dimension $(p \times p)$ where $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda p$ are eigen values of $\boldsymbol{\Sigma}$. Then TV $= \mathrm{Tr}(\boldsymbol{\Sigma}) = \lambda_1 + \lambda_2 + \ldots + \lambda p$, CD $=$ GV $= |\boldsymbol{\Sigma}| = \lambda_1 \cdot \lambda_2 \cdot \ldots \cdot \lambda_p$ and VV $= \lambda_1^2 + \lambda_2^2 + \ldots + \lambda_p^2$.

Robust distance of MVV is a robust distance generated by the estimator MVV. This estimator is obtained from a criterion which minimizes the vector variance (VV). Just like MCD, MVV estimator possesses high break down point and of affine equivariant. Another good property of MVV is that MVV possesses smaller time complexity than MCD, i.e. $O(p^2)$. Herwindiati et al [13] discussed the MVV algorithm in detailed fashion. A brief description of the algorithm is as follow: given n random samples $\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_n$ of dimension a p-variate. Suppose T_{MVV} and $\boldsymbol{C}_{\mathrm{MVV}}$ are MVV estimators for location parameters and covariance matrix. Both estimators are defined based on a set H consists of $h = [\frac{n+p+1}{2}]$ data which gives covariance matrix $\boldsymbol{C}_{\mathrm{MVV}}$ of minimum $\mathrm{Tr}(C^2_{\mathrm{MVV}})$ among all possible h data. Then,

$$T_{\mathrm{MVV}} = \frac{1}{h} \sum_{\vec{X}_i \in \mathrm{H}} \vec{X}_i \tag{1}$$

$$\boldsymbol{C}_{\mathrm{MVV}} = \frac{1}{h} \sum_{\vec{X}_i \in \mathrm{H}} (\vec{X}_i - T_{\mathrm{MVV}})(\vec{X}_i - T_{\mathrm{MVV}})^t \tag{2}$$

The one of the goodness estimator properties is the affine equivariant; which is property is not influenced by affine transformation. MVV estimator is affine equivariant estimator. Consider random samples $\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_n$ of random vectors $\vec{X}$ of location parameter $T \in \mathbb{R}^p$ and scale parameter C in the space of $p \times p$ symmetric matrices. Suppose $\boldsymbol{X}$ defines an $n \times p$ matrix where the k-th row is $\vec{X}_k^t$. A location estimator $T_n(\mathbf{X}) \in \mathbb{R}^p$ is said to have the affine equivariant property if for every vector $\vec{b} \in \mathbb{R}^p$ and every nonsingular $p \times p$ matrix $\boldsymbol{A}$ the condition

$$T_{\mathrm{MVV}}(\boldsymbol{A}\boldsymbol{X} + \vec{b}) = \frac{1}{h} \sum_{\vec{X}_i \in \mathrm{H}} (\boldsymbol{A}\vec{X}_i + \vec{b}) = \boldsymbol{A}T_{\mathrm{MVV}} + \vec{b} \tag{3}$$

$$\begin{aligned} \boldsymbol{C}_{\mathrm{MVV}}(\mathrm{AX} + \vec{b}) &= \frac{1}{h} \sum_{\vec{X}_i \in \mathrm{H}} (\boldsymbol{A}\vec{X}_i - \boldsymbol{A}T_{\mathrm{MVV}})(\boldsymbol{A}\vec{X}_i - \boldsymbol{A}T_{\mathrm{MVV}})^t \\ &= \boldsymbol{A}\boldsymbol{C}_{\mathrm{MVV}}\boldsymbol{A}^t \end{aligned} \tag{4}$$

4. The MVV Robustness Measure

The breakdown point is a quantitative measure to describe the concept of robustness. Rousseeuw and Leroy [20] defined from the context of a sample that breakdown point is the smallest fraction of data which causes the value of estimator to be infinity when the value of all data in the fraction are changed to be infinity. The concept of breakdown point is highly related to the concept of estimator bias. Consider $T_n(\boldsymbol{X})$ and $\boldsymbol{C}_n(\boldsymbol{X})$ are mean vector and covariance matrix of sample. Suppose the estimator $T_n(\boldsymbol{X})$ becomes $T_n(\boldsymbol{X}^*)$ if the value of m data are changed. Rousseeuw and Leroy [20] define breakdown point $\varepsilon_n^*(T, \vec{X})$, for sample of size n as follows

$$\text{bias}(m, T, \vec{X}) = \sup_{\boldsymbol{X}^*} \|T_n(\boldsymbol{X}^*) - T_n(\boldsymbol{X})\| \tag{5}$$

$$\varepsilon_n^*(T, \vec{X}) = \min\{\frac{m}{n} | \text{bias}(m, T, \vec{X}) \text{ infinite}\} \tag{6}$$

MVV is high breakdown point. The following figures: Figure 2, Figure 3 and Figure 4 are illustrated the breakdown point FSA, FMCD and MVV.

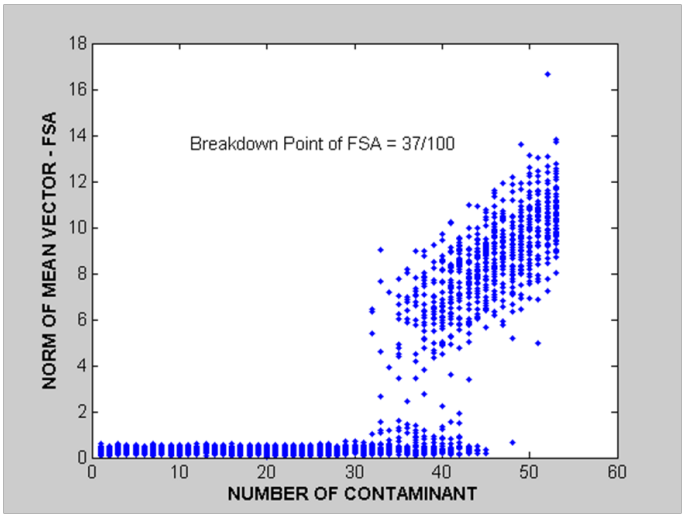

Figure 2. The Simulation Breakdown Point of FSA

Founded on figures, the smallest fraction causing the MVV estimator to be infinity is 52/100. It means that the fraction is broken toward infinity if the contaminants present almost an half of data. The fraction is not smaller than FSA and FMCD.

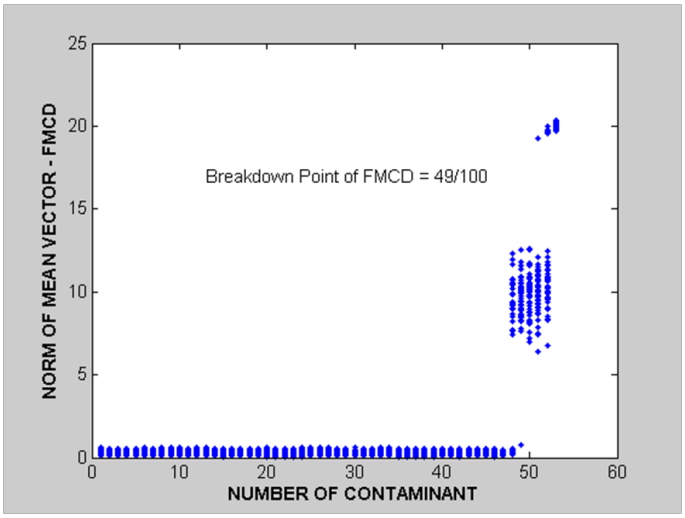

FIGURE 3. The Simulation Breakdown Point of FMCD

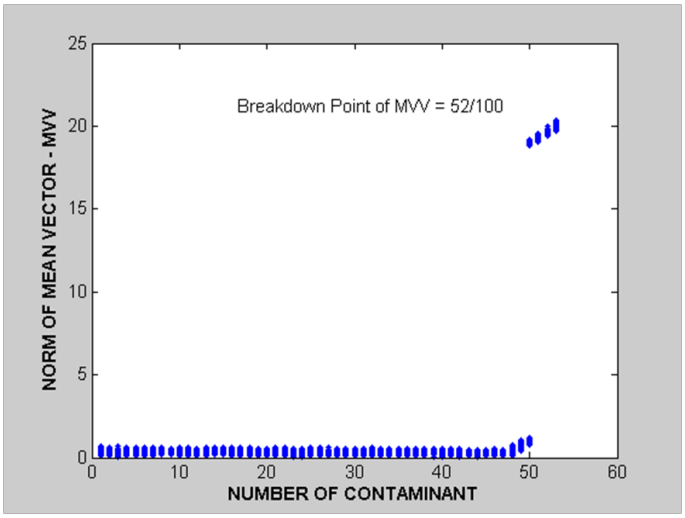

FIGURE 4. The Simulation Breakdown Point of MVV

5. The Influence of Outlyingness Observations on MVV Estimators

The observation is called influential if its deletion would cause major change in estimate and diagnosis statistics, Kotz and Johnson [15]. The classical estimator is very sensitive on to influential observations. The occurrence of one or more outliers shifts the mean vector toward outliers and the covariance matrix becomes to be inflated. The deletion of an influential observation or outliers can significantly

change to the estimator. The estimator is said not sensitive if there is no significant change due to removal of outlier.

There are many ways to measure the sensitivity; this paper comes up bringing simple discussion, both on computations and the theoretical distribution. Consider data set $\mathrm{X}_n = \{\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_n\}$ of p-variate normal distribution, the scatter matrix of sample $\boldsymbol{A}$ is

$$\boldsymbol{A} = \sum_{j=1}^{n} (\vec{X}_j - \vec{\bar{X}})(\vec{X}_j - \vec{\bar{X}})^t \tag{7}$$

where $\vec{\bar{X}} = \frac{1}{n}\sum_{j=1}^{n} \vec{X}_j$, $\vec{\bar{X}}$ the sample's mean vector. From equation [7], the scatter matrix $\boldsymbol{A}$ is of Wishart distribution with parameter $\boldsymbol{\Sigma}$ and degree of freedom $n-1$, written as $\boldsymbol{A} \sim W_p(\boldsymbol{\Sigma}, n-1)$, $\boldsymbol{A}$ is independent of $\vec{\bar{X}}$ (see Wilks [26]).

Define $\boldsymbol{A}_{-i}$ the scatter matrix removing the i^{th} observation, say the i^{th} observation is an outlying observation. The scatter matrix $\boldsymbol{A}_{-i}$ by is formulated as

$$\boldsymbol{A}_{-i} = \sum_{i \neq j=1}^{n} (\vec{X}_j - \vec{\bar{X}}_{-i})(\vec{X}_j - \vec{\bar{X}}_{-i})^t \tag{8}$$

with $\vec{\bar{X}}_{-i} = \frac{1}{n-1}\sum_{i \neq j=1}^{n} \vec{X}_j$. The scatter matrix $\boldsymbol{A}_{-i}$ is of Wishart distribution with parameter $\boldsymbol{\Sigma}$ and the degree of freedom $n-2$, $\boldsymbol{A}_{-i} \sim W_p(\boldsymbol{\Sigma}, n-2)$.

The ratio of scatter matrix as the consequence of removal the i^{th} observation is given by

$$\boldsymbol{R}_i = \frac{|\boldsymbol{A}_i|}{|\boldsymbol{A}|} = \frac{\left|\boldsymbol{A} - \vec{b}_i\vec{b}_i^t\right|}{|\boldsymbol{A}|} = \frac{\left|1 - \vec{b}_i^t\boldsymbol{A}^{-1}\vec{b}_i\right||\boldsymbol{A}|}{|\boldsymbol{A}|} = 1 - \vec{b}_i^t\boldsymbol{A}^{-1}\vec{b}_i$$

$$\vec{b}_i = \sqrt{\frac{n}{n-1}}\left(\vec{X}_i - \vec{\bar{X}}\right)$$

$\vec{b}_i$ is of Normal distribution $N_p\left(\vec{0}, \boldsymbol{\Sigma}\right)$ and independent of $\boldsymbol{A}_i$ (see Mardia et. al. [18]). $\boldsymbol{R}_i$ can be shown of distribuion $beta\left(\frac{n-p-1}{2}, \frac{p}{2}\right)$. The proof was discussed in Mardia et. al. [18] and rewritten in Appendix completed with surveying the probability density function which was skipped by Mardia et. al. as an exercise. Since the distribution of $1 - \vec{b}_i^t\boldsymbol{A}^{-1}\vec{b}_i$ is $\boldsymbol{beta}\left(\frac{\boldsymbol{n-p-1}}{\boldsymbol{2}}, \frac{\boldsymbol{p}}{\boldsymbol{2}}\right)$ then $\boldsymbol{\vec{b}_i^t A^{-1}\vec{b}_i \sim beta}\left(\frac{\boldsymbol{p}}{\boldsymbol{2}}, \frac{\boldsymbol{n-p-1}}{\boldsymbol{2}}\right)$.

The distribution of classical approach is well known, it is different with the robust approach. The distribution of robust is not easy to be composed. Usually we have to do the simulation approach to get the distribution. In the section will be discussed the sensitivity and the approximated distribution of robust approach.

Let data set $\mathrm{X} = \{\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_n\}$ of p-variate observations. If observations taken from it a subset $\mathrm{H} \subseteq \mathrm{X}$ consist of h data points, then $\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_h$ are random sample of size h and of distribution $N_p(\vec{\mu}, \boldsymbol{\Sigma})$, h assumed as $h = [\frac{n+p+1}{2}]$.

The location and scale estimator can be computed as,

$$\vec{\bar{X}}^R = \frac{1}{h} \sum_{\vec{X}_i \in \mathrm{H}} \vec{X}_i \tag{9}$$

$$\boldsymbol{A}^R = \sum_{\vec{X}_i \in \mathrm{H}} (\vec{X}_i - \vec{\bar{X}}^R)(\vec{X}_i - \vec{\bar{X}}^R)^t \tag{10}$$

Based on limit central theory, if $\vec{X}_1, \vec{X}_2, \ldots, \vec{X}_n \sim N_p(\vec{\mu}, \boldsymbol{\Sigma})$ then the distribution of $\boldsymbol{S}^R$ can be approximated by

$$mc^{-1}\boldsymbol{S}^R \sim W_p(\boldsymbol{\Sigma}, m) \tag{11}$$

(see Hardin and Rocke [11]). It means that

$$\boldsymbol{A}^R = \sum_{\vec{X}_i \in \mathrm{H}} (\vec{X}_i - \vec{\bar{X}}^R)(\vec{X}_i - \vec{\bar{X}}^R)^t = c^{-1}\frac{\boldsymbol{S}^R}{m} \tag{12}$$

If $c = 1$ then $\boldsymbol{A}^R \sim \frac{1}{m} W_p(\boldsymbol{\Sigma}, m)$.
The ratio of robust scatter matrix because of an outlier deletion

$$\boldsymbol{R}_i^R = \frac{|\boldsymbol{A}_{-i}^R|}{|\boldsymbol{A}^R|}.$$

To draw an analogy between the clasical approach to the robust approach, the distribution of $\boldsymbol{R}_i^R$ approximated by

$$\boldsymbol{R}_i^R \sim \mathrm{beta}(\frac{n-p-2}{2}, \frac{p}{2}).$$

The estimator is said to be insensitive to an outlier when

$$\boldsymbol{R}_i^R > \mathrm{beta}(\frac{n-p-2}{2}, \frac{p}{2}).$$

For the illustration of sensitivity of classical and MVV robust measure, let the multivariate data having size $n = 100$, $p = 10$ data contain $k = 2$ outliers which are far from a bulk of data. The sensitivities of two methods, the classic and robust MVV method, are measured by using the ratio $\boldsymbol{R}_{-k}$. Table 1.1 has contents of ratio.

TABLE 1. The Ratio of $\boldsymbol{R}_{-k}$

The Value	**The Method: Classical**	**Robust**
$\boldsymbol{R}_{-k}$	0.123119	0.916667
Cut Off	0.958491	0.853018
Sensitivity to outliers	**Very sensitive***	**Insensitive**

The ratio of classical method is very small. It means that the removing of outliers causes the serious problem on the classical estimator. The value of estimator is very sensitive to outliers. It can be seen in the Table 1.1, the estimator

becomes to be inflated when the outliers 'present' on the data set. The contrary with the classical sensitivity, the MVV estimator is not sensitive from 'presenting' or removing outliers.

6. Experiment and Analysis

In this section is discussed the robust performance of MVV in two CBIR experiments on flower images by colors and textures.

The first experiment is carried out to find out whether the robust MVV distance able to differ species of flowers of 'almost the same colors'. For this purpose, it is chosen 3 species of flowers of **red colour** and **almost red** of different texture, i.e. Red Hibiscus, Red Rose and Potentilla Napalensis as the following.

Figure 5. The Images of Red Flower (a) Red Hibiscus (b) Red Rose and (c) Potentilla Napalensis

The first stage of CBIR process is to build an image database by extracting the characteristics of whole database. The extraction of characteristic is carried out to catch characteristic of color through the color moment method and the characteristic of texture through the wavelet transform feature. The color moment method yields characteristic vector of dimension 14×1, meanwhile the wavelet transform feature method yields characteristic vector of texture of 7×1. The images which are used are of format bitmap **(.bmp)** of maximum resolution $m \times n$ pixels. Having characteristic extraction stage, all resulted vectors are stored in the database.

The next stage of CBIR is to process the image retrieval, which is started from entering query image. The images then processed to find out the value of color characteristic extraction and the texture characteristic. The characteristic extraction process on the query image yields two characteristic vectors, the dimension of these vectors will be the same with characteristic vectors in the images database. Similarity of the result of retrieval process is measured by using the robust MVV distance.

The retrieval process still needs a perception of human. Machines just can be used to search the relevant images. The performance of retrieval process is represented by the precision of the appearance of relevant retrieved images. The precision is defined as the probability of 'success' to find retrieved image matching query image.

In the first experiment, the image of red hibiscus, red rose and potentilla napalensis will be 'searched' from a database red flower image for 30 times. In the each retrieval process will be appeared 5 relevant images. The precision of retrieval process will be listed in Table 1.2.

TABLE 2. The Precision of The First Experiment

The Species of Flower	**Mean of Precision: Mahalanobis Distance**	**MVV Robust Distance**
Red Rose	90%	92%
Red Hisbiscus	72%	96%
Potentilla Napalensis	66%	94%
Grand Mean of Precision	**76%**	**94%**

In the second experiment, it will be checked whether MVV is able to measure the similarity of flowers in a better way. By choosing 5 species of flowers of color 'tend to red', and 'tend to purple', three of the same species as previous experiment, other species are Linum Narbonense and Geranium Psilostemon as follow

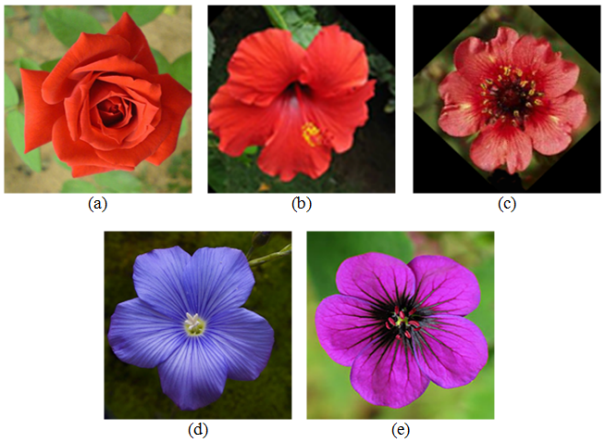

FIGURE 6. The Images of Flower (a) Red Hibiscus (b) Red Rose and (c) Potentilla Napalensis, (d) Linum Narbonense (e) Geranium Psilostemon

The experiments above tell us that MVV robust distance is more effective than Mahalanobis distance for the CBIR similarity measure. The summaries of retrieval precision of two methods are appeared in Table 1.2 and Table 3. The surprised results that are found in the second experiment MVV is still robust to effect a 'fusion' of red color and purple color.

Table 3. The Precision of the Second Experiment

The Species of Flower	**Mean of Precision: Mahalanobis Distance**	**MVV Robust Distance**
Red Rose	80%	72.5%
Red Hisbiscus	72.5%	82.5%
Potentilla Napalensis	42.5%	92.5%
Linum Narbonense	55%	87.5%
Geranium Psilostemon	42.5%	77.5%
Grand Mean of Presicion	**50%**	**82.5%**

7. Remark

From the experiment, it is unveiled that the robust MVV distance is a reliable measure to measure the propinquity of characteristics of an image query to the data base image.

References

[1] Alt, F. B. and Smith, N. D., "Multivariate Process Control", *Handbook of Statistics* **7** (1988), 333 - 351.

[2] Anderson, T.W., *An Introduction to Multivariate Statistical Analysisa*, Second Edition, John Wiley, New-York, 1984.

[3] Angiuli, F. and Pizzuti, C., "Outlier Mining and Large High Dimensional Data Sets", *IEEE Transaction on Knowledge and Data Engineering* **17(2)** (2005), 203 - 215.

[4] Barnett, V. and Lewis, T., *Outliers in Statistical Data*, Second Edition, John Wiley, New-York, 1984.

[5] Becker, C. dan Gather, U., "The Masking Breakdown Point of Multivariate Outlier Identification Rules", *Journal of American Statistical Association* **94** (1999), 947 - 955.

[6] Billor, N., Hadi, A.S. and Velleman, P.F., "BACON: blocked adaptive computationally efficient outlier nominators", *J. Computational Statistics and Data Analysis* **34** (2000), 279 - 298.

[7] Derquenne, C., "Outlier Detection Before Running Statistical Methods", *J. Siam* **34(2)** (1992), 323 - 326.

[8] Djauhari, M.A., "Improved Monitoring of Multivariate Process Variability", *Journal of Quality Technology* **37(1)** (2005), 32 - 39.

[9] Hadi A.S., "Identifying multivariate outlier in multivariate data", *Journal of Royal Statistical Society B* **53(3)** (1992), 761 - 771.

[10] Hardin, J dan Rocke, D. M., "The Distribution of Robust Distance", (2002), http://www.cipic.ucdavis.edu/ dmrocke/preprints.html

[11] Hardin, J dan Rocke, D. M., "The Distribution of Robust Distance", *J. of Computation and Graphical Statistics* **14** (2005), 928 - 946.

[12] Hawkins, D.M., "Feasible Solution Algorithm for the Minimum Covariance Determinant Estimator in Multivariate Data", *J. Computational Statistics and Data Analysis* **17** (1994), 197 - 210.

[13] Herwindiati, D.E., Djauhari, M.A. and Mashuri, M., "Robust Multivariate Outlier Labeling", *J. Communication in Statistics Simulation And Computation* **36(6)** (2007)

[14] Huber, P. J., "Robust Estimation of Location Parameter", *Annals of Mathematical Statistics* **35** (1964), 73 - 101.
[15] Kotz S dan Johnson N.L., *Encyclopedia of Statistical Sciences*, John Wiley, New York, 1985.
[16] Irwin, J.O., "On a Criterion for the Rejection of Outlying Observations", *J. Biometrics* **17(3/4)** (1925), 238 - 250.
[17] Long, F., Zhang, H. and Feng, D.D., *Multimedia Information Retrieval and Management*, Spinger, Berlin, 2003.
[18] Mardia, K. V., Kent, J. T., and Bibby, J. M., *Multivariate Analysis*, Academic Press, London, 1979.
[19] Rousseeuw, P.J., "Multivariate Estimation with High Breakdown Point", Paper appered in Grossman W., Pflug G., Vincze I. dan Wertz W., editors, *Mathematical Statistics and Applications* **B** (1985), 283 - 297. D. Reidel Publishing Company.
[20] Rousseeuw, P.J and Leroy A. M., *Robust Regression and Outlier Detection*, John Wiley, New-York, 1987.
[21] Rousseeuw, P.J. dan Van Zomeren, B.C., "Unmasking Multivariate Outliers and Leverage Points", *Journal of American Statistical Association* **85(441)** (1990), 633 - 639.
[22] Rousseeuw, P.J. and van Driessen, K., "A Fast Algorithm for The Minimum Covariance Determinant Estimator", *J. Technometrics* **41** (1999), 212 - 223.
[23] Thompson, W. R., "On a Criterion for the Rejection of Observation and the Distribution of the Ratio of Deviation to Sample Standard Deviation", *J. The Annals of Mathematical Statistics* **6(4)** (1935), 214 - 219.
[24] Werner, M., *Identification of Multivariate Outliers in Large Data Sets*, PhD. Thesis, University of Colorado at Denver, 2003.
[25] Wilks, S. S., *Finite-Dimensional Vector Spaces*, Springer-Verlag, New-York, 1974
[26] Wilks, S. S., "Multivariate Statistical Outliers", *J. Sankya A* **25** (1963), 407 - 426.
[27] Ye, N., Borror, C.M., and Parmar, D., "Scalable Chi-Square Distance versus Conventional Statistical Distance for Process Monitoring with Uncorrelated Data Variables", *Quality and Reliability Engineering International* **19** (2003), 505 - 515.

APPENDIX

By applying the Equation A.2.3m in Mardia et. al [18], we have:

$$\boldsymbol{R}_i = \frac{|\boldsymbol{A}_i|}{|\boldsymbol{A}|} = \frac{|\boldsymbol{A}_i|}{\left|\boldsymbol{A}_i + \vec{b}_i\vec{b}_i^t\right|} = \frac{|\boldsymbol{A}_i|}{\left|1 + \vec{b}_i^t\boldsymbol{A}_i^{-1}\vec{b}_i\right||\boldsymbol{A}_i|} = \frac{1}{1 + \vec{b}_i^t\boldsymbol{A}_i^{-1}\vec{b}_i}$$

Theorem 1.1. *$\vec{b}_i^t\boldsymbol{A}_i^{-1}\vec{b}_i$ has the same distribution with $\frac{p}{n-p-1}W$ where $W \sim F_{p,n-p-1}$.*

Proof 1.1. *Recall that $\vec{b}_i$ and $\boldsymbol{A}_{-i}$ are independently distributed as $N_p\left(\vec{0}, \boldsymbol{\Sigma}\right)$ and $W_p(\boldsymbol{\Sigma}, n-2)$, respectively. Set $\vec{d} = \Sigma^{-\frac{1}{2}}\vec{b}_i$ and $M = \Sigma^{-\frac{1}{2}}A_i\Sigma^{-\frac{1}{2}}$, then $\vec{d} \sim N_p\left(\vec{0}, \boldsymbol{I}\right)$, $M \sim W_p(\boldsymbol{I}, n-2)$ and independent each other. We see that $\alpha = (n-2)\vec{d}^t M^{-1}\vec{d}$ has Hotelling T^2 distribution with parameter p and $n-2$ by Definition 3.5.1 of Mardia et. al* [18] *and by Theorem 3.5.2 of the book, $\alpha = (n-2)\vec{d}^t M^{-1}\vec{d}$ has the same distrbution with $\frac{(n-2)p}{(n-2)-p+1}W$ where $W \sim F_{p,(n-2)-p+1}$. But*

$$\begin{aligned}
(n-2)\vec{d}^t M^{-1}\vec{d} &= (n-2)\left(\Sigma^{-\frac{1}{2}}\vec{b}_i\right)^t\left(\Sigma^{-\frac{1}{2}}A_i\Sigma^{-\frac{1}{2}}\right)^{-1}\left(\Sigma^{-\frac{1}{2}}\vec{b}_i\right)\\
&= (n-2)\vec{b}_i^t\left(\Sigma^{-\frac{1}{2}}\right)^t\left(\Sigma^{-\frac{1}{2}}\right)^{-1}A_i^{-1}\left(\Sigma^{-\frac{1}{2}}\right)^{-1}\Sigma^{-\frac{1}{2}\vec{b}_i}\\
&= (n-2)\vec{b}_i^t\Sigma^{-\frac{1}{2}}\left(\Sigma^{-\frac{1}{2}}\right)^{-1}A_i^{-1}\left(\Sigma^{-\frac{1}{2}}\right)^{-1}\Sigma^{-\frac{1}{2}\vec{b}_i}\\
&= (n-2)\vec{b}_i^t A_i^{-1}\vec{b}_i
\end{aligned}$$

so, $\vec{b}_i^t A_i^{-1}\vec{b}_i$ has the same distribution with $\frac{p}{(n-2)-p+1}W$ where $W \sim F_{p,(n-2)-p+1}$. Since $(n-2)-p+1 = n-p-1$ then we can say that $\vec{b}_i^t A_i^{-1}\vec{b}_i$ has the same distribution with $\frac{p}{n-p-1}W$ where $W \sim F_{p,n-p-1}$.

Theorem 1.2. *If $Y = \frac{1}{1+\frac{r}{s}W}$ where $W \sim F_{r,s}$ then $Y \sim Beta\left(\frac{s}{2}, \frac{r}{2}\right)$.*

Proof 1.2. *From $Y = \frac{1}{1+\frac{r}{s}W}$ we have that $W = \left(\frac{1}{Y} - 1\right)\frac{s}{r}$. Since the probability density function (pdf) of W is*

$$\begin{aligned}
f(w) &= \frac{\Gamma\left(\frac{r+s}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{s}{2}\right)}\left(\frac{r}{s}\right)^{\frac{r}{2}}\frac{w^{\frac{r}{2}-1}}{\left(1+\frac{r}{s}w\right)^{\frac{r+s}{2}}}, 0 < w < \infty\\
&= 0, \textit{ elsewhere}
\end{aligned}$$

then the pdf of Y *is*

$$\begin{aligned}
h(y) &= f\left(w(y)\right)\left|\frac{df}{dy}\right| \\
&= \frac{\Gamma\left(\frac{r+s}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{s}{2}\right)}\left(\frac{r}{s}\right)^{\frac{r}{2}}\frac{\left(\left(\frac{1}{y}-1\right)\frac{s}{r}\right)^{\frac{r}{2}-1}}{\left(1+\frac{r}{s}\left(\frac{1}{y}-1\right)\frac{s}{r}\right)^{\frac{r+s}{2}}}\left|-\frac{1}{y^2}\frac{s}{r}\right| \\
&= \frac{\Gamma\left(\frac{r+s}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{s}{2}\right)}\left(\frac{s}{r}\right)^{-\frac{r}{2}+1}\frac{\left(\frac{1}{y}-1\right)^{\frac{r}{2}-1}\left(\frac{s}{r}\right)^{\frac{r}{2}-1}}{\left(\frac{1}{y}\right)^{\frac{r+s}{2}}}\frac{1}{y^2} \\
&= \frac{\Gamma\left(\frac{r+s}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{s}{2}\right)}\frac{\left(\frac{1}{y}-1\right)^{\frac{r}{2}-1}}{\left(\frac{1}{y}\right)^{\frac{r+s}{2}}}\frac{1}{y^2} \\
&= \frac{\Gamma\left(\frac{r+s}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{s}{2}\right)}\frac{\left(\frac{1-y}{y}\right)^{\frac{r}{2}-1}}{(y)^{-\frac{r+s}{2}}}y^{-2} \\
&= \frac{\Gamma\left(\frac{r+s}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{s}{2}\right)}(1-y)^{\frac{r}{2}-1}y^{-\frac{r}{2}+1+\frac{r+s}{2}-2} \\
&= \frac{\Gamma\left(\frac{r+s}{2}\right)}{\Gamma\left(\frac{r}{2}\right)\Gamma\left(\frac{s}{2}\right)}(1-y)^{\frac{r}{2}-1}y^{\frac{s}{2}-1}, 0<y<1 \\
&= 0, \textit{ elsewhere}
\end{aligned}$$

we see that $h(y)$ *is the pdf of* $Beta(\frac{s}{2},\frac{r}{2})$.

Corollary 1.3. $R_i \sim Beta(\frac{n-p-1}{2},\frac{p}{2})$

Proof 1.3. *From Theorem [1.1] and Theorem [1.2] we have that* $\frac{1}{1+\vec{b}_i^t\boldsymbol{A}_i^{-1}\vec{b}_i} \sim Beta(\frac{n-p-1}{2},\frac{p}{2})$ *by replace* r *with* p *and* s *with* $n-p-1$.