http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 13 (6): 1001-1013, 2014
ISSN 1812-5638 / DOI: 10.3923/itj.2014.1001.1013

Visual Clothing Search by Shape and Style

Jyh-Jong Tsay, Chi-Hsiang Lin and Tsung-Yueh Lai
Department of Computing and Information Science, School of Information Engineering,
College of Physical and Engineering Science, Yangzhou University, Canada

Abstract: With a rapidly increasing use of the Internet, there is a large amount of digital image information, more and more picture information is applied to the commercial website. How to retrieve the digital image in database efficiently has been an significant issue. Besides described by words, pictures of clothes products are often attached. Traditionally, we tag each image with several keywords, so that users can search keywords to find products. However, the key words may contain no clue of what they are looking for. Clothes products often contain many characteristics which are hard to describe in regular keywords, such as texture, shape, or any other subtle details. To tackle this issue, we develop the image-based visual clothing search system, which enables users to retrieve clothing images. In this study, we provide the visual search with shape using the "Shape context" to be the shape descriptor so that one can search for similar shapes of clothes efficiently. We develop the system with two steps. First, appropriate feature description from clothes image is extracted . Second, retrieval strategies of clothes image in database is built. We experiment on an image database with several kinds of clothes type and the experiment shows that our approach is more helpful for ranking clothes product by shape and style.

Key words: Clothing retrieval, visual search, shape context, CBIR

INTRODUCTION

As the amount of digital multimedia content grows on the Internet, online shopping plays the indispensable role in our life. The digital image has become increasingly popular in recent years. Therefore, it is necessary to develop a searching engine utilizing content-based image retrieval in addition to the conventional text-based searching engine. How to help consumers find what they want successfully is always a challenge for online shopping websites. In such searching engine based on images, users are able to search with "pictures" instead of "key words". The system returns users images similar to the pictures of their inputs and users may get a desirable result. The system resembles the "Visual Internet Object Search (VIOS)" system developed by Chang and Tsay (2010) and it provides functions of searching texture, neck types and color of clothes products for users. Nevertheless, in order to obtain a more accurate query result, merely depending on these properties is still not enough. In searching clothes, the most important things are shape and style, then, the secondary ones are texture and color. To overcome the problem that shape and style is not included in the VIOS system, we have to develop a shape-based retrieval function. In this study, we will provide a better approach to solve such problems. By using "Shape context" as the feature descriptor, images are represented with collections of "visual words". Moreover, we adopt text search approaches to process "visual words" to retrieve similar images.

Traditionally, image features analysis uses color, shape and texture. The whole statistical histogram matching as a kind of relatively ripe method has been applied to a lot of general-purpose image retrieval system. However, they are mainly for rigid objects and cannot be directly applied to soft objects such as clothing. Although, humans can easily detect different styles and parts of clothes, these differences are inherently difficult for a computer. Clothing is soft and often distorted by texture, distortion, wrinkling and shading (Fig. 1), which influence image identification. Furthermore, clothing with

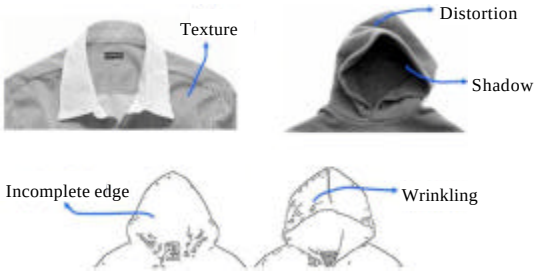


Fig. 1: Examples of influence affecting image identification

Corresponding Author: Jyh-Jong Tsay, Department of Computing and Information Science, School of Information Engineering, College of Physical and Engineering Science, Yangzhou University, Canada

similar styles can have different patterns, textures and decorations. All of these make clothing search more complex than image search of rigid objects. In this study, we aim to develop robust, efficient and effective image retrieval techniques to handle above problems.

RELATED WORK

Content-Based Image Retrieval (CBIR) basically makes use of visual features extracted from an image to describe it (Tseng *et al*., 2009). One of the main difficulties in clothes image searching is the variability of appearance of a target object, such as folding, self-occlusion and distortion. Moreover, efficient searching in large-scale image database is also a challenge for computer systems. This study will focus on the problems about the variability of clothes appearance and build a framework for clothes retrieval based on the shape. Users can search similar kinds of clothes through the framework which increases the precision by overcoming some interfering factors, such as the changes in illumination, folding and partial occlusion. In the following related work, we will introduce the VIOS system, visual vocabulary, text-retrieval method and a garment visual search based on shape matching.

Visual vocabulary and text-retrieval method: In typical Content-Based Image Retrieval (CBIR) system, it is always important to select an appropriate representation for documents. Thus, the retrievals from images and the classification of images in database are significant issues.

As we can see from some previous studies. Sivic and Zisserman (2003) proposed an object retrieval approach using the techniques of text mining. They apply the concept of Bag-Of-Words (BOW) and Bag-Of-Feature (BOF) (Csurka *et al*., 2004) to image retrieval. Bag-Of-Words (BOW) refers to a set of similar feature descriptors which are represented by a word providing visual analogy. A word providing visual analogy is termed "visual word". To represent images with visual words, the following are some steps they take. First, they cluster the feature descriptors in all images and construct a visual vocabulary. Then, each local feature descriptor extracted from images is mapped to a specific visual word existing in the visual vocabulary. In this way, images are viewed as documents constituted by visual words, which are centroids of clusters formed by clustering local feature descriptors.

By using technology Term Frequency-Inverse Document Frequency (TF-IDF), an image is represented by a vector consisting of weighted visual words. After the vector of query is compared with that of images in the database, the ranking can be produced according to the similarity between vectors.

Visual internet object search (VIOS) system: Traditionally, we tag each image with several keywords, so, that users can search keywords of images and products. However, this method is more suitable for users who have already understood the searched object clearly. Therefore, to tackle the issue that users might want to search objects with images instead of keywords Chang and Tsay (2010) developed an image-based visual clothing retrieval system, which extracts and uses the features of clothing images to find objects that are difficult to describe by text or annotations. They integrated techniques of image processing and information retrieval into develop the retrieval system and then converted low-level visual features into high-level semantic concepts (human perception of visual semantics).

In the study, they provide a search interface called "Visual Internet Object Search (VIOS)" system, which allows users to find out the clothes in which they are interested from images similar to user's query. Overall, they presented a framework with clothes image search system which provides functions of searching ROI, neck style and color of clothes. The proposed framework is mainly based on the techniques of text mining and image processing. They introduce two feature descriptors, SIFT feature descriptor and ART-based feature descriptor. Meanwhile, they apply the concept of visual word to similar SIFT feature descriptors and collect visual words to use in soft-based weighting technology. Finally, instances of VIOS system search results of image according to ROI similarity, neck style similarity and color similarity are shown in Fig. 2-4.

Garment visual search based on shape context: Tseng *et al*. (2009) presented a matching method to search garment images by using the shape feature. Their solution on clothes retrieval consists of three phases: Segmentation, feature extraction and shape matching. First, they apply the segmentation approach to eliminate some part of texture, such as wrinkles and shading which might be the noise and reduce the accuracy. However, the segmentation process might bring some side-effect like lost information of hems and sewing lines (Fig. 5). After the segmentation stage, the original image is represented by a sketchy and coarse segmentation line graph. Second, in the step of extracting features, they take sample points from the edge elements on the shape. These points can be on internal or external contours. A sampling instance is shown in Fig. 6. Third, they use "shape context" feature

Fig. 2: VIOS system: Search results of the query image by ROI similarity

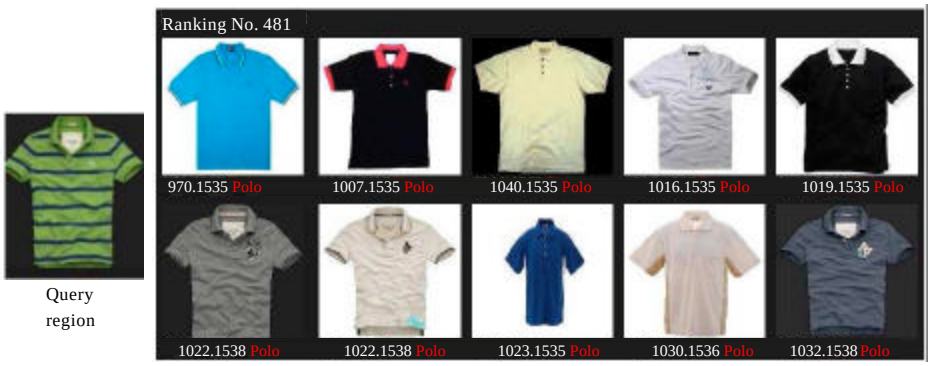


Fig. 3: VIOS system: Search results of the query image by neck

Fig. 4: VIOS system: Search results of the query image by color

Fig. 5: A sample of a segmented image

descriptor to get a vector for each image and calculate the cost of matching two points by χ^2 test statistic. Finally, the correspondence cost can be calculated by average distance of matching.

However, compared with our search for index, the matching method takes a large amount of time in the matching step and the accuracy is not quite desirable. We will compare the retrieval result of our method to that of their method in the experimental stage.

Fig. 6: An example of sampling points

OVERVIEW OF FRAMEWORK

In this section, we present a general overview of our method, in which the shape retrieval framework includes two parts: Off-line processing and on-line processing (Fig. 7). In terms of off-line processing, we will show how to build it which includes shape feature extraction and building index. First, the characteristics of clothes images collected in the database will be presented. Second, the extraction of foreground and segmentation of clothes will be introduced. Third, the feature extraction and indexing are brought up. On the other hand, in terms of on-line processing, the processing of ranking will be introduced.

Database images: The purpose of adding the option of shape in the system is that users can search clothes with the image condition. We use different shapes of clothes to set up different data sets. In order to make the database more suitable for the large-scale online shopping websites, we will continuously update the database. For keeping diversity, clothes images are collected from different clothes websites, such as online auction, NIKE, Giordano, Hang Ten, Uniqlo, Lativ, etc., several common styles of clothes are included, such as cake dresses, camises, dangle dresses, girl vests, hat coats, long sleeve T-shirts, etc. The 1064 items of clothes are classified into 14 categories according to their shape. Each category consists of at least 50 images. The collected images are all in plain backgrounds and in JPEG format. In the experiment, we use these categories to evaluate the result

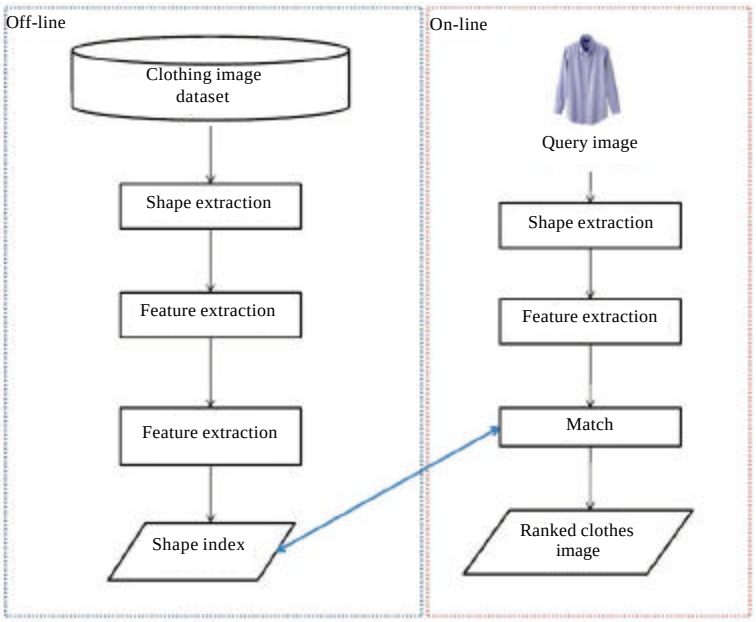


Fig. 7: The flowchart of presented framework

of our framework on shape and style and we combine the functions of shape and style with neck, color and ROI from the VIOS system. Some trials will be performed and the result will be evaluated in order to improve the system.

Foreground extraction: In order to obtain the shape of clothes, foreground/background segmentation is used to distinguish the contour of clothes. We adopt the approach based on the method "GrabCut" proposed by Rother *et al.* (2004). The method can always obtain the accurate segmentation of objects from background. The "GrabCut" method is based on graph cut algorithm which is proposed by Boycov and Jolly (2001) and the basic idea of graph cut is turning images into monochrome image of gray level and making foreground/background segmentations by using the energy minimization between two pixels. Two enhancements to the graph cut mechanism have been made: "iterative estimation" and "incomplete labeling", which together allow a considerably reduced degree of user interaction for a given quality of result.

In this way, we treat the extracted foreground as the shape of the clothes, leaving contours of the edge in the image. Compared with traditional segmentation which leaves more internal texture, the "GrabCut" method will leave less internal texture. Instead of leaving only contours, the method leaves a little internal texture probably. The method makes use of internal texture left in shape context to describe clothes. The shape context descriptor will contain some internal points of some clothes, so that users can find the most similar images at the retrieval stage. In a nutshell, when we use the "GrabCut" method, most texture, wrinkles, shading, decorations and inner patterns are eliminated (Fig. 8).

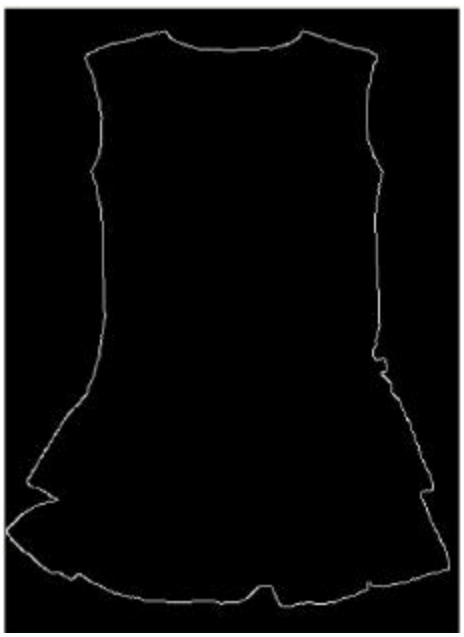


Fig. 8: Sampled shape edge of the clothes

Feature extraction and indexing: The system uses the properties extracted from the shape of the whole image after image segmentation is processed to retrieve similar images in database. A rich feature descriptor which produces useful feature description is very important. In this study we use shape context (Belongie and Malik, 2000; Mori *et al*., 2001; Belongie *et al*., 2002) to extract the local visual features. Shape context digitizes the object curves by using the log-polar plane. Then, we cluster the extracted local feature descriptor. The centroids of each cluster refer to the representative features, which are also called visual words. The images in the database are represented with visual word vectors. We can weight the visual word vector of each image according to the number of the visual words appearing in every image from database. Eventually, the weighted vector is applied to indexing, so, that the speed and precision of searching will be improved in the process of retrieval. TF-IDF is used to weight the vector and it is modified to advance the accuracy of ranking. Finally, we create the forward table to store the weighted vector and build inverted table to speed up the search.

Ranking images: In the ranking stage, when the user enters a query image, we first extract the shape of the query image and then each keypoint is also extracted and described by the shape descriptor in the feature extraction stage. After the query vector is calculated, we measure the similarity between the query vector and the vector of each image in the database. The clothes images will be ranked from high to low similarity.

SHAPE FEATURE AND INDEXING

An effective framework and an efficient feature descriptor are important to retrieve images. The performance of retrieval results is always influenced by feature descriptions. This chapter discusses the retrieval system and shape feature descriptors. The indexing method and ranking mechanism are also included.

Local feature extraction: In this study, we introduce how to generalize the concept from text-base search to non-textual information. In the search of clothes images, each image can be represented by a set of local descriptors. We apply "Shape Context" feature descriptors which

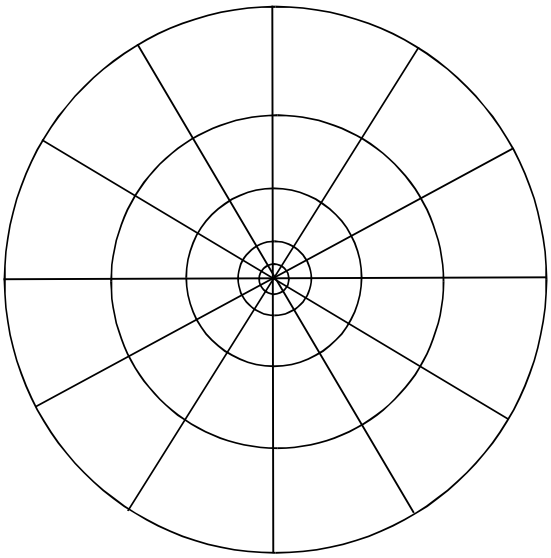
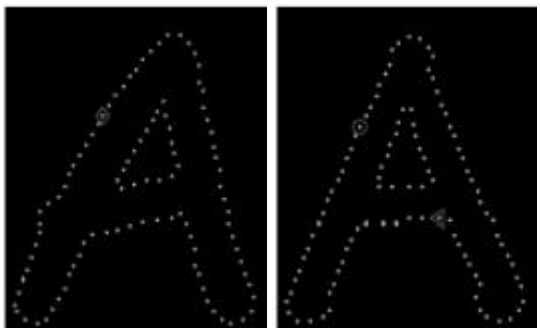


Fig. 9: Diagram of log-polar histogram bins

are most commonly used to describe the shape of objects (Mori *et al*., 2005; Tangelder and Veltkamp, 2004). The shape context is described by the relative position of each point nearby in the image. Shape context descriptors make the comparison of objects more accurate, because they describe the shape of the outside and the texture of parts inside. They also digitize the object curves by using the log-polar plane. After the "GrabCut" method is used in the shape extraction stage, the shape context analysis begins by taking N samples from the edge elements on the shape. We use the Harris corner detection to get the best N corners under a minimum distance gap. If the number of corners is less than the demanded sampling number, some remaining points will be selected randomly to meet the demanded number as closely as possible. In our framework, we set the 100 points to be the sample number N in each image.

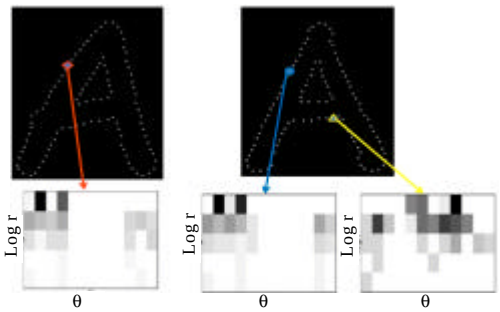
Shape context: Shape context is a kind of descriptor which is usually used to describe the shape of objects. According to Belongie and Malik (2000), the basic idea of shape context is illustrated in Fig. 9.

A shape is represented by a discrete set of points sampled from the internal of external contours on the image shape. In order to produce the shape context of an object, the edge pixels are derived by using the Grabcut to estimate foreground edge of cloth image. Thus, the shape context analysis is performed by taking n sample points from a curve. Figure 10 a and b are two examples of taking sample points.



(a) (b)

Fig. 10 (a-b): Sampled edge points of two shapes



(a) (b)

Fig. 11(a-b): Example shape contexts reference samples

So, a shape of an object can be represented by a discrete set $P = \{p_1,..., p_n\}$, $p_i \in R^2$ of n points. The set of vectors is originated from a point to all other sample points on a shape. These n-1 vectors express the configuration of the entire shape relative to the reference point. One way to capture the information is to observe the distribution of the relative positions of the remaining n-1 points in a spatial histogram. To be more specific, for a point p_i on the shape, the way of calculating its coarse histogram h_i is based on the relative coordinates of the remaining n-1 points:

$$h_i(k) = \#\{q \neq p_i : (q - p_i) \in bin(k)\} \quad (1)$$

This histogram is defined as the shape context of p_i. We use bins that are uniform in log-polar space, making the descriptor more sensitive to positions of nearby sample points than to those of points farther away. Figure 11a and b illustrate that sample points are represented by the coarse histogram. The shape contexts for reference samples marked by the symbols -◇ ○ △. Note the visual similarity of the shape context for -◇ ○, which were computed for relatively similar points on the two shapes. By contrast, the shape context for △ is quite different. In our framework, we apply the log-polar histogram

that is composed of 5 log-space radius for log r and 12 identical angles for θ. Therefore, after described by shape contexts, each keypoint is represented by a 60-bins vector.

Visual vocabulary construction: As mentioned earlier, after extracting Shape context feature descriptors from each image in database, we quantize these descriptor vectors into centroids, which provides visual analogy of words called the "visual word". In order to apply the approach of text retrieval to image search, we build a visual vocabulary according to those visual words. This step utilizes K-means algorithm to classify all shape context feature descriptors in hyper space when k clusters get k centroids. Each centroid is called a visual word, which is a 60-dimensional vector. All visual words build up a visual vocabulary, which plays an essential role in the image retrieval.

Typically, it is an important issue to know how to choose the number of cluster centers to maximize retrieval performance on ground truth data of manually labeled objects. In the experiment, we observe that what affects quality of an individual visual word is the result of clustering. There are two situations that the number of clusters might vary: First, if the number of cluster centers is relatively large, visual vocabulary will become less generalizable and may cause more noise and extra overhead processing. On the other hand, if the number of cluster centers is relatively small, it may lack the discriminative power since two key points may be assigned into the same cluster even if they are not similar to each other.

The trade-off between discrimination and generalization motivates the study of visual vocabulary size. Our survey shows that previous works used a distinctively different range of vocabulary sizes, resulting in difficulties in interpreting their findings. For instance, Zhang *et al*. (2007) adopted 1000, Elsayad *et al*. (2010) adopted 1000-4000, Sivic and Zisserman (2008) adopted 6000-10000, etc. In our study, we experiment on vocabularies with various numbers of visual words. Empirically, 600-800 visual words are found to be the most appropriate parameter in our retrieval system.

Typically, feature selection is an important technique in text categorization for reducing the vocabulary size and the feature dimension. Traditionally, in text categorization, words with small Document Frequency (DF) are significant since rarely-used words are usually informative in category prediction. In contrast, the visual words that occurs in almost all images, as a stop word, is not special and significant for retrieval. In order to filter out non-informative visual words, we generate a stop word list with a threshold after calculation Document Frequency (DF) of each visual word. DF is the number of images (documents) in which a visual word appears (frequency). Generally speaking, common visual words occur in many places of an image. In this case, visual words with DF>90% are removed from visual vocabulary.

Forward file and inverted table: In the indexing step, we will use a visual words vector to represent each image, at first, we extract the feature descriptors and each feature descriptor is quantized by mapping to the ID of the nearest visual word. Suppose there are k words in a visual vocabulary and each clothes image is represented by a vector T:

$$T = [t_1, ..., t_i, ...t_k] \quad (2)$$

where, t_i denotes term i (the ith visual word), which is the weighted bin of frequency in the images. The detail of weighting scheme for the visual word is introduced. At the retrieval stage, each image includes some particular visual words. According to Eq. 2, each weighted vector for images are generated. For visual index of shape context, we build a "forward table" which stores the weighted vector in each images (Fig. 12).

For ranking retrieved image, forward table is used to calculate cosine similarity of the weighting vector. There are three properties included by the forward table: Image name, image ID and weighted vector of the image. Moreover, we use the forward table to generate the inverted index table to speed up searching. The inverted index table stores lists of image ID numbers per visual word (Fig. 13).

Feature weighting and ranking mechanism: The Term Frequency-Inverse Document Frequency (TF-IDF) scheme is the modern weighting technology often used in information retrieval and text mining. It is a statistic approach to measure the significance of the word in "specific document" and "all documents". The importance of the visual word is directly proportional to the frequency they occur in the document; whereas, it is inversely proportional to the frequency they occur in all documents. For weighting the visual word, we use a k-dimensional vector $T = [t_1,...,t_i, ...,t_k]$ to represent an image. Each bin t_i within the TF-IDF formula is:

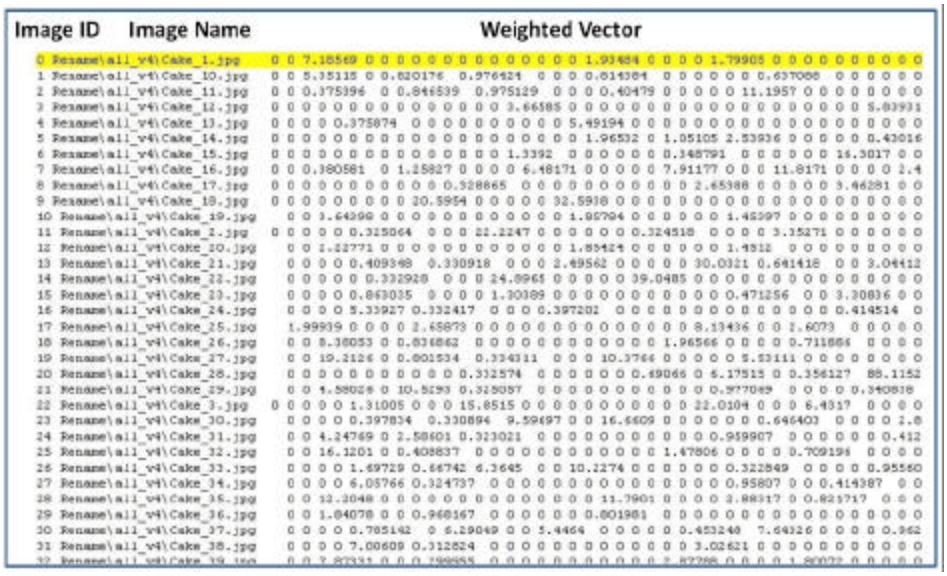


Fig. 12: Structure of forward table

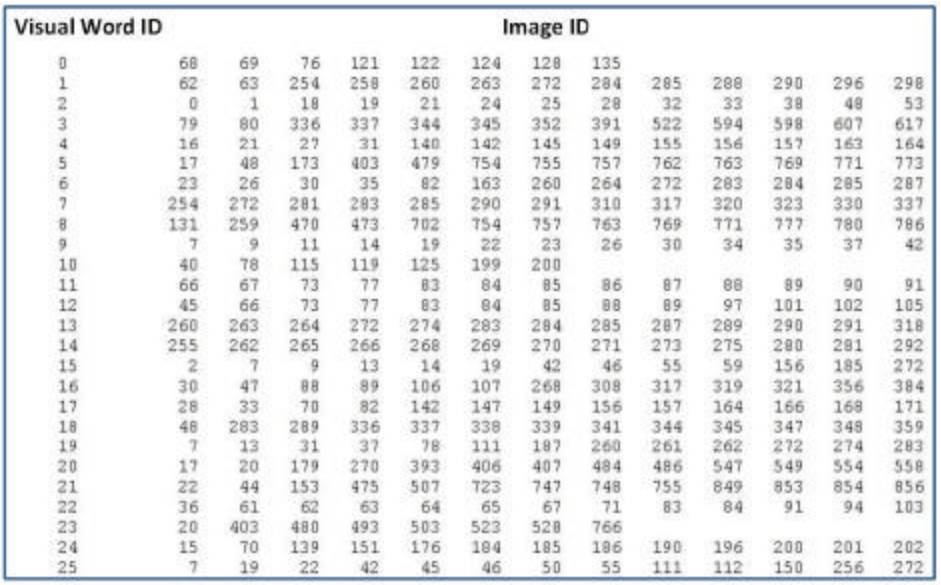


Fig. 13: Structure of inverted index table

$$t_i = \frac{n_{id}}{n_d} \log \frac{N}{N_i} \quad (3)$$

where, n_d denotes the total number of visual words in the image d, n_{id} denotes the number of occurrences of visual word i in the image d, N_i is the number of images containing visual word i and N is the number of all images in the database. There are two concepts included in the weighting approach: First, n_{id}/n_d is the frequency of the visual word, which illustrates the importance of the visual word in the image. Second, N/N_i is used to adjust the weight of visual word frequency. That illustrates the particularity of the visual word in all images. In conclusion, TF-IDF weighting scheme is used in measure of the importance of the visual words. At the retrieval step, images are ranked according to the normalized scalar product. The similarity formula is:

$$Sim(T_q, T_d) = \frac{T_q^t T_d}{\|T_q\|_2 \|T_d\|_2} \quad (4)$$

$$\|T\|_2 = \sqrt{T^T T} \quad (5)$$

where, T_q denotes the shape context vector of the query image and T_d denotes the vector of each image in the database. When ranking images, we measure the similarity between the query vector and the vector of each image in the database by the cosine of the included angle. Equation of Bober (2001) is the L^2-norm of the vector. The similarity will normalized between 0 and 1. In our framework, users input a clothes image and the shape of the clothes will be extracted and become some shape context features. Furthermore, the specified visual words are generated and form the query vector.

All these weighting schemes perform the nearest neighbor search in the vocabulary in the sense that each keypoint is mapped to the most similar visual word. However, assigning the specific visual word to each keypoint only according to its nearest neighbor is not an optimal situation. One possible reason is that two similar points may be clustered into different clusters when the size of visual vocabulary increases. Moreover, simply counting the term frequency may not be optimal as well. In order to tackle the aforementioned problems, Jiang *et al.* (2007) proposed a novel soft-weighting method to assess the significance of a visual word to an image. For each keypoint in an image, instead of searching only for the nearest visual word, Tsay *et al.* (2011) presented several weighting methods of selecting the Nearest-R nearest visual words. The following is the way how to build Nearest-R based weighting scheme. Suppose we have a visual vocabulary of k visual words and we use a k-dimensional vector $T = [t_1,...,t_i,...,t_k]$ in which each component t_i represents the weight of a visual word i in an image below:

$$t_i = \sum_{r=1}^{R} \sum_{j=1}^{M_r} \frac{1}{2^{r-1}} \text{sim}(j,i) \quad (6)$$

where, M_r represents the number of keypoints whose rth nearest neighbor is visual word i. The measure sim (j,i) represents the similarity between keypoint j and visual word i. Notice that in Eq. 6, the contribution of a keypoint depends on its similarity to word i weighted by $1/2^{r-1}$. In other words, the weighted vector is affected by not only the nearest centroids but also any other visual words nearby, showing that the closer the visual word is, the more sensitive the weight becomes.

Base on the Eq. 3 and the Eq. 6, the modified method of counting keypoints and weighting into two equations is as follows: the first equation is called "Soft-IDF" weighting method:

$$t_i = \sum_{r=1}^{R} \sum_{j=1}^{M_r} \frac{1}{2^{r-1}} \text{sim}(j,i) \bullet \log \frac{N}{N_i} \quad (7)$$

The second equation is called "Soft-RS-IDF" weighting method:

$$t_i = \sum_{r=1}^{R} \sum_{j=1}^{M_r} \frac{\text{sim}(j,i)}{\sum_r^R \text{sim}(j,i)} . \log \frac{N}{N_i} \quad (8)$$

By using the proposed soft-based weighing scheme, we expect to address the fundamental drawbacks of the conventional weighing schemes and we will experiment and evaluate this kinds of weighting methods in the retrieval stage.

EXPERIMENT AND RESULT

The images we use to evaluate performance and a method of evaluating the result of retrieval called Mean Average Precision (MAP) will be introduced in the chapter. In this chapter, we present four parts of evaluation. Firstly, we evaluate the effect of different weighting methods. Secondly, we calculate Mean Average Precision (MAP) based on different visual vocabulary sizes that significantly affects the accuracy of the retrieval system. Thirdly, after compared to the matching method of the garment visual search, our method is proved to be more accurate and faster. Lastly, we evaluate the MAP based on different soft-weighting parameters, which are Nearest-R nearest visual words. Besides, with the hit rate, it is proved that our method works well in garment matching. Moreover, the result of retrieval will be presented and the performance will be discussed and evaluated.

Queries: In our database, all categories of 1064 pieces of clothes are contained. In the experiment, a query set includes all images of the 1064 pieces of clothes instead of selecting images in the database randomly, so that the result of evaluation will present the most complete performance of the system.

Performance and retrieval example: We evaluate the influence of different weighting method in terms of Mean Average Precision (MAP). The average precision is the most commonly used measure to retrieve performance.

The MAP equation is as follows:

$$\text{MAP}(Q) = \frac{1}{G} \sum_{i=1}^{G} \frac{i}{p} \quad (9)$$

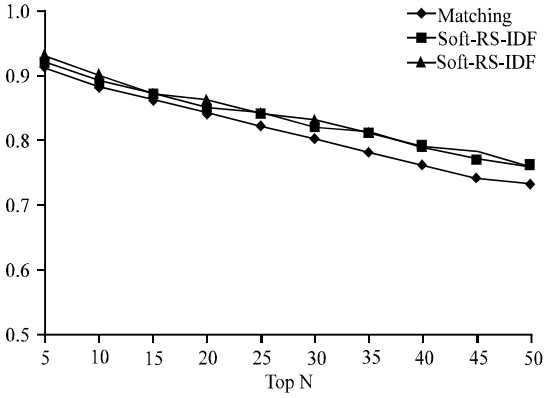
where, Q denotes the query, G denotes the number of relevant images in the database and p denotes the ranking position of relevant image. For example, consider a query that has four relevant images which are retrieved from ranks 1, 2, 4 and 7. The actual precision of each rank is 1, 1, 0.75 and 0.57, respectively. Thus, the average precision over all relevant images for this query is 0.83, which is also the MAP.

Evaluate different weighting methods: Three methods of weighting are mentioned in section 4.5, that is, TF-IDF, Soft-IDF and Soft-RS-IDF. We use the MAP of top N ranked images to measure our retrieval system and apply these weighting methods to different vocabulary sizes. In soft-based weighting we select the Nearest-4 nearest visual words as the soft-parameter. The retrieval performance is evaluated in the test of MAP for top 50 images. The result is presented in Fig. 14.

The result of the performance of top 50 ranked images in the entire clothes database is observed from the above. We see that the weighting method Soft-RS-IDF performs better than any other methods, especially when the vocabulary size reaches a certain amount. It is indicated that the soft-based weighting method referencing visual words nearby the keypoint works better than the method referencing only the closest visual word to the keypoint. Therefore, Soft-RS-IDF performs better than TF-IDF in such a vocabulary size and it is the reason why we adopt Soft-RS-IDF.

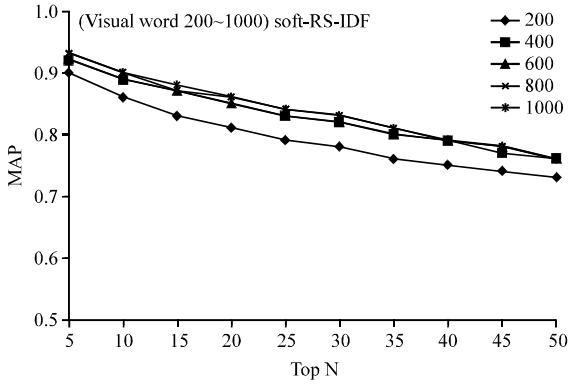
Evaluate different visual vocabulary sizes: Typically, the number of visual words is chosen empirically to maximize retrieval performance on a manually labelled image or ground truth data. The k-means algorithm use iteration to find the result of minimum cost. These centers of clusters are described in shape context. In our retrieval system, the number of visual words (clusters) influences the result of retrieval significantly. The Soft-RS-IDF is confirmed to be the best weighting method. Next, the impact of different vocabulary sizes is examined. By using Soft-RS-IDF weighting method, we evaluate the MAP for top 50 ranked images with different numbers of visual words and show the result in Fig. 15.

From the above, the effect of the number of visual words is analyzed with the mean-average precision in our retrieval system. In our experiment, we build different visual vocabulary sizes, such as 200, 400, 600, 800, 1000 visual words. The weighting method we use is Soft-RS-IDF and the number of visual words is considered to be the most appropriate from 600-800. Because when visual vocabulary size becomes too small, such as 200 visual words, the same visual word might appear in images that are quite different. In this way, the visual words would no longer contrast to one another. On the other hand, when the vocabulary size includes from 800-1000 visual words, however, we could not see this improvement on MAP. As a result, we choose 800 visual words as our vocabulary size in our system.



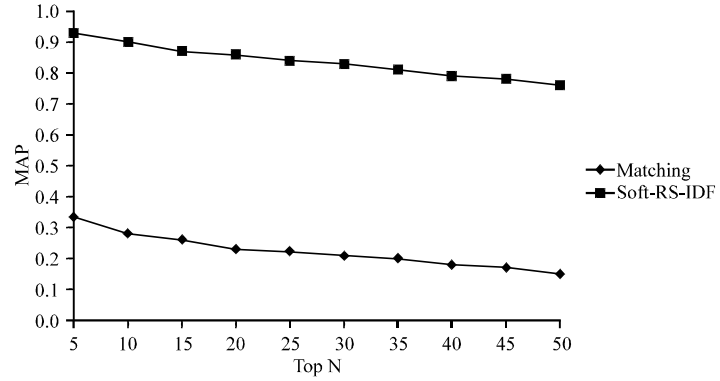
	5	10	15	20	25	30	35	40	45	50
TF-IDF	0.91	0.88	0.86	0.84	0.82	0.80		0.76	0.74	0.73
Soft-IDF	0.92	0.89	0.87	0.85	0.84	0.82	0.81	0.79	0.77	0.76
Soft-RS-IDF	0.93	0.90	0.87	0.86	0.84	0.83	0.81	0.79	0.78	0.76

Fig. 14: MAP of different weighting method based on 800 visual words



	5	10	15	20	25	30	35	40	45	50
200	0.90	0.86	0.83	0.81	0.79	0.78	0.76	0.75	0.74	0.75
400	0.92	0.89	0.86	0.84	0.82	0.81	0.79	0.78	0.76	0.75
600	0.93	0.90	0.87	0.85	0.83	0.82	0.80	0.79	0.78	0.76
800	0.93	0.90	0.87	0.86	0.84	0.83	0.81	0.79	0.78	0.76
1000	0.93	0.90	0.87	0.86	0.84	0.83	0.81	0.79	0.78	0.76

Fig. 15: Number of visual words and MAP for top 50 ranked images



	5	10	15	20	25	30	35	40	45	50
Matching	0.334	0.28	0.26	0.23	0.222	0.21	0.20	0.18	0.17	0.15
Soft-RS-IDF index	0.93	0.90	0.87	0.86	0.86	0.83	0.81	0.79	0.78	0.76

Fig. 16: Matching method and Indexing method with soft-RS-IDF weighting

Evaluate different retrieval method: Tseng *et al.* (2009) presented a matching method to search garment images by using the shape feature. In our experiment, 124 pieces of clothes from 14 categories are selected to be query images. It takes a lot of time to match as the paper indicates. To get retrieval with the matching method, there are a few steps to follow: First, the images in the query images and database will be segmented. Second, 100 points described by shape context will be extracted from each query image and image from the database. Third, the method of x^2 is used to calculate the cost of the query image and the image compared, so that we can get the similarity between them. Finally, we rank the similarity from high to low and use MAP to measure the result.

The MAP of the matching method and TF-RS-IDF indexing method is shown Fig. 16. The total 124 query images takes 10418 sec to process and in average each takes 84 sec in C^{++} program. However, in indexing method we propose, most time is spent on preprocess. We spend

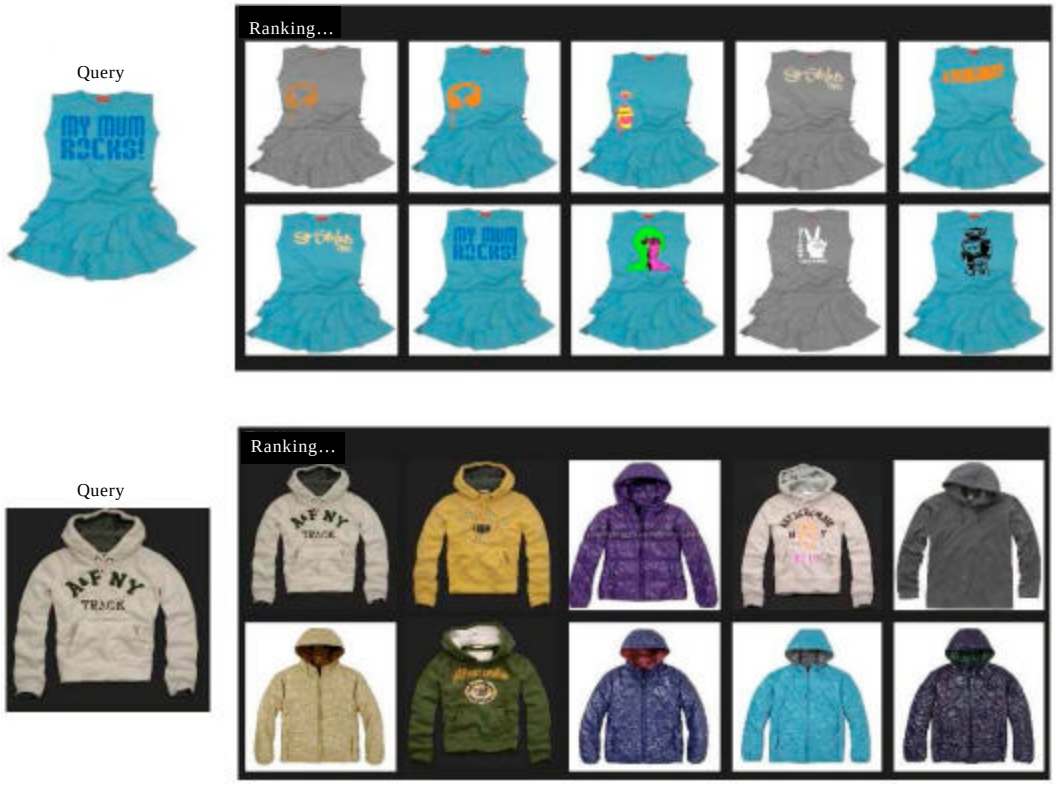


Fig. 17: The good retrieval result of the cake dress and sweatshirt

1.33 sec to extract the shape of query image and average 8.5 sec on the stage of retrieval according to forward table and inverted table. Therefore, our method shortens the time between entering the image and getting the result of searching.

Examples of shape query and retrieval: Figure 17 show the examples of good retrieval results from the top 10 images retrieved from the database. Using shape context based bag-of-visual-words is beneficial for the accuracy of searching.

CONCLUSION

The study proposes a framework for clothing image search by shape. The shape descriptor called "Shape context" is used to describe shape features. The main idea of shape context is described by the relative position of each point nearby in the image. Meanwhile, we utilize the concept of visual word by clustering the similar shape context descriptors and collect visual words with soft-based weighting technology. As evidenced from the experiments, soft-based weighting is a better weighting method for visual words than the commonly used TF-IDF. A framework is proposed for users to retrieve the clothes products by shape on the Internet auctions or online shopping. It can help consumers to search products more efficiently. In the experiments, our framework is also compared with the matching method for garment visual search proposed by Tsen *et al*., making the accuracy and processing speed raised.

REFERENCES

Belongie, S. and J. Malik, 2000. Matching with shape contexts. Proceedings of the IEEE Workshop on Content-based Access of Image and Video Libraries, June 12, 2000, Hilton Head Island, pp: 20-26.

Belongie, S., J. Malik and J. Puzicha, 2002. Shape matching and object recognition using shape contexts. IEEE Trans. Patt. Anal. Mach. Intell., 24: 509-522.

Bober, M., 2001. MPEG-7 visual shape descriptors. IEEE Trans. Circuits Syst. Video Technol., 11: 716-719.

Boycov, Y.Y. and M.P. Jolly, 2001. Interactive graph cuts for optimal boundary and region segmentation of objects in N-D images. Proceedings of the 8th IEEE International Conference on Computer Vision, July 7-14, 2001, Vancouver, BC., pp: 105-112.

Chang, K.C. and J.J. Tsay, 2010. A study of visual clothing search. CSIE Department, Chung Cheng University, ROC, 2010.

Csurka, G., C.R. Dance, L. Fan, J. Williamowski and C. Bray, 2004. Visual categorization with bags of keypoints. Proceedings of the IEEE Workshop on Statistical Learning in Computer Vision, May 2004, Meylan, France, pp: 1-16.

Elsayad, I., J. Martinet, T. Urruty and C. Djeraba, 2010. A new spatial weighting scheme for bag-of-visual-words. Proceedings of the International Workshop on Content-Based Multimedia Indexing, June 23-25, 2010, Grenoble, pp: 1-6.

Jiang, Y.G., C.W. Ngo and J. Yang, 2007. Towards optimal bag-of-features for object categorization and semantic video retrieval. Proceedings of the 6th ACM International Conference on Image and Video Retrieval, July 9-11, 2007, ACM, New York, pp: 494-501.

Mori, G., S. Belongie and J. Malik, 2001. Shape contexts enable efficient retrieval of similar shapes. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Volume 1, December 8-14, 2001, IEEE Computer Society, pp: 723-730.

Mori, G., S. Belongie and J. Malik, 2005. Efficient shape matching using shape context. IEEE Trans. Pattern Anal. Mach. Intell., 27: 1832-1837.

Rother, C., V. Kolmogorov and A. Blake, 2004. GrabCut: Interactive foreground extraction using iterated graph cuts. ACM Trans. Graphics, 23: 309-314.

Sivic, J. and A. Zisserman, 2003. Video Google: A Text Retrieval Approach to Object Matching in Videos. IEEE Computer Society, Washington, Pages: 1470.

Sivic, J. and A. Zisserman, 2008. Efficient visual search for objects in videos. Proc. IEEE, 96: 548-566.

Tangelder, J.W.H. and R.C. Veltkamp, 2004. A survey of content based 3D shape retrieval methods. Multimedia Tools Appl., 39: 441-471.

Tsay, J.J., C.H. Lin, C.H. Tseng and K.C. Chang, 2011. On visual clothing search. Proceedings of the International Conference on Technologies and Applications of Artificial Intelligence, November 11-13, 2011, IEEE Computer Society, pp: 206-211.

Tseng, C.H., S.S. Hung, J.J. Tsay and D. Tsaih, 2009. An efficient garment visual search based on shape context. WSEAS Trans. Comput., 8: 1195-1204.

Zhang, J., M. Marszalek, S. Lazebnik and C. Schmid, 2007. Local features and kernels for classification of texture and object categories: A comprehensive study. Int. J. Comput. Vision, 73: 213-238.