Measuring Clothing Image Similarity with Bundled Features

<table>
<thead>
<tr>
<th>Journal:</th>
<th>International Journal of Clothing Science and Technology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID:</td>
<td>IJCST-Sep-2012-0058</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>Research Paper</td>
</tr>
<tr>
<td>Keywords:</td>
<td>Clothing image similarity, Local word frequency, SIFTs distance matrix, Bundled feature</td>
</tr>
</tbody>
</table>
Measuring Clothing Image Similarity with Bundled Features

Abstract

Purpose - Clothing retrieval is very useful to help the clients to efficiently search out the apparel they want. Currently, the mainstream clothing retrieval methods are attribute semantics based, which however are inconvenient for common clients. The purpose of this paper is to provide an easy-to-operate apparels retrieval mode with our novel approach of clothing image similarity measurement.

Design/methodology/approach - We measure the similarity between two clothing images by computing the weighted similarities between their bundled features. Each bundled feature consists of the point features (SIFT) which are further quantified into local visual words in a maximally stable extremal region (MSER). We weight the importance of bundled features by the precision of SIFT quantification and local word frequency that reflects the frequency of the common visual words appeared in two bundled features. The bundled features similarity is computed from two aspects: 1) local word frequency; 2) SIFTs distance matrix that records the distances between every two SIFTS in a bundled feature.

Findings - Local word frequencies improves the recognition between two bundled features with the same common visual words but different local word frequency. SIFTs distance matrix has the merits of scale invariance and rotation invariance. Experimental results show that our approach works well in the situations with large clothing deformation, background exchange and part hidden, etc. And the similarity measurement of Weight+Bundled+LWF+SDM is the best.

Originality/value - This paper presents an apparel retrieval mode based on local visual features, and presents a new algorithm for bundled feature matching and apparel similarity measurement.

Keywords Clothing image similarity, Local word frequency, SIFTs distance matrix, Bundled feature

Paper type Research paper

1. Introduction

Clothing online sales are increasing rapidly and have been doubled in the last five years. How to help the clients to efficiently search out the apparel they want is a key problem influencing the core competiveness of the clothing e-commerce provider. The key to this problem is providing an intelligent clothing retrieval system.

Most of the existing clothing retrieval systems are clothing attribute semantics based (Haiping et al., 2002). For searching out appropriate apparel, the client should firstly input the exact attribute semantic words of the apparel. However, due to the cognitive difference, it is very difficult to unify attribute semantics for the apparel in a large variety of styles, which affects the effectiveness of the retrieval results. We have presented a clothing similarity measurement method based on TLAC to improve the retrieval performance (Zheng Liu et al., 2012). Moreover, describing the attribute semantics for proper apparel usually calls for expertise of clothing, which makes it inconvenient for common clients.

Image based retrieval technologies are increasingly applied in a large number of areas including online clothing sales due to their intuition and convenience. When the client input a clothing image which is close to
his/her demands, image based retrieval systems are expected to intelligently output a list of clothing images which have visual similarity with the inputting image. Such a easy-to-operate retrieval mode is much more attractive than attribute semantics based retrieval mode.

The key technology for image based retrieval systems is computing the visual similarity between two images, which however is not easy to be resolved. As soft objects, the shape of apparel can be largely different under various physical conditions, even for same apparel, as shown in Figure 1. Non-rigid deformation, shadow, and part hidden may exist in the clothing images. Additional to apparel, information irrelevant to apparel such as human bodies and backgrounds may also exist in the clothing images. Therefore, how to unify the visual features of clothing images becomes a key problem of image based clothing retrieval.

Figure 1: Examples of clothing images on web.

The existence of non-rigid clothing deformation, shadow, part hidden and uncertain background determines that the commonly used image features for rigid objects, such as color, silhouette, texture, etc (Worring et al., 2000; Kota et al., 2012; Zhang et al., 2004; Datta et al., 2008) may lose the discriminative power, and they are not ideal for the clothing images. We have noticed that large non-rigid deformation of clothing mostly occurs in the small regions near human body joints, while it is relatively small in the regions in-between every two adjacent joints and such regions occupies the most area of the clothing image. Therefore, the most of the local regions have relative stable visual appearance. Thus we choose proper local image features as the basic features for similarity calculation. Specially, the SIFT feature (Lowe et al., 2004) has the merits of scale, rotation and lighting changing invariance, which is very suitable for computing the similarity between clothing images.

SIFT features based image retrieval methods compute the similarity between every two images by matching their local SIFT descriptors. Zhao et al. (2007) proposes a one-to-one symmetric matching algorithm and D. Xu et al. (2008) employs multi-level spatial matching. However, tens of thousands of SIFT features can be extracted from a common image and each SIFT feature is usually described in a high dimensional space (128 dimensions in our approach), which makes the matching computationally expensive. For efficiency, J. Sivic et al. (2003) treats the image as a set of visual words each of which is obtained by clustering the SIFT features with k-mean clustering. Their method is inspired by text retrieval. Philbin et al. (2007; 2008) and Nister et al. 2006 use hierarchical k-means clustering which has better clustering effect to build a vocabulary, and then the SIFT features are quantified to build a scalable vocabulary tree, which improves the retrieval efficiency. However the quantization step limits the discriminative power, then geometric verification (Philbin et al., 2007; H. Jegou et al., 2008)
becomes an important post-processing step to improve discrimination ability.

Zhong et al. (2009) and Zhipeng et al. (2010) bundle the local visual words in each maximally stable extremal region (MSER) (Matas et al. 2004) and add geometric constraint to the bundled feature. Experiments show that the bundled feature is much more discriminative than an individual SIFT feature. Zhong et al. (2009) compute the similarity of two bundled features between different images by calculating the number of common visual words and comparing the ordering relationships among SIFT features. However, the weak geometric constraint is sensitive to large rotations. Zhipeng et al. (2010) improve the bundled feature with an affine invariant geometric constraint, they use every two SIFT features’ positions and the center of SIFT features to form a triangle, then the normalized triangles’ areas are used to form a matrix that is further served as the geometric constraint in evaluating the similarity between two bundled features. This geometric constraint is robust to image scale and large rotation transformation. However, the flip invariant they proposed is incorrect e.g., after getting the SIFT features in one image, then flip it, the new SIFT features computed from the flipped image might have the same respective positions as their previous, however the feature descriptors will be different, and the corresponding local visual words will be changed. Moreover, it is possible that two SIFTs-formed triangles are similar in area but largely different in shape; therefore their geometric constraint has its inadequacy in describing the spatial relationship among SIFT features. What’s more, each common visual word may relative to different number of SIFT features, i.e. the frequency of each visual word may different in a bundled feature. Such difference cannot be distinguished by their method, because of their indiscriminate treatment of the common visual words with different frequencies.

Our approach is also bundled features based. For a query clothing image, the similarity between it and a random image in the database is computed as the weighted sum of similarity between bundled features. Since flipping the clothing images are really meaningless for online sales, and seldom would the clothing images be flipped on online clothing shopping webs, our approach treats the clothing images regardless of its flipping. Our approach has a close algorithm structure with Zhong et al. (2009) and Zhipeng et al. (2010), however our approach improves the detailed methods in setting geometric constraints and computing the similarity between bundled features. Our approach has the following merits.

1. A LWF-based method is proposed for computing the similarity between bundled features. The frequencies of each common visual word appeared in bundled features and their difference are used to compute the similarity, which overcomes the problem of low discrimination between two bundled features with the same common visual words but different local word frequency.

2. A SDM-based method is proposed for computing the geometric similarity between bundled features. SDM has the merits of scale invariance and rotation invariance and well describes the spatial relationship among SIFT features, which overcomes the problem of low discrimination between two bundled features that have similar SIFTs-formed triangle areas but different shapes.

3. A method based on LWF and SDM is proposed for setting the matching between bundled features on different clothing images and computing their similarity. The whole clothing image similarity is computed by a weighted sum of bundled feature similarities. The weight is computed by the precision of SIFT features quantification and the local word frequency.

The rest of the paper is organized as follows. The bundled features of clothing image is extracted in section 2. Bundled feature matching based on LWF and SDM is proposed in section 3. Similarity between clothing images is
2. Bundled features

Bundled features are formed by bundling the SIFT features in each MSER. Supposing that $S = \{s_j\}$ are SIFT features on a clothing image, $R = \{r_i\}$ are the region features (MSER). Thus, the bundled features $B = \{b_i\}$ can be defined as:

$$b_i = \{s_j \mid s_j \prec r, s_j \in S\}$$

where $s_j \prec r$ means $s_j$ falls inside region $r$. $b_i$ is discarded when the region $r$ does not include any SIFT point. The overall characteristic model of a clothing image is defined by the set of bundled features as $C = \{b_i\}$. Figure 2 shows the forming process of bundled features, Figure 2(a) is the original clothing image, Figure 2(b) shows the clothing image SIFT features in red points, Figure 2(c) shows the extracted region features (MSER) indicated by ellipses, Figure 2(d) shows the bundled features motivated by SIFT and MSER.

![Figure 2: Forming process of bundled features.](image)

3. Bundled features matching based on LWF and SDM

Let $p = \{p_i\}$ and $q = \{q_i\}$ be two bundled features in two clothing images ($C_a$ and $C_b$), $p_i$ and $q_i$ are SIFT features. The similarity between $p$ and $q$ can be measured through two aspects: (1) the relevancy of visual words between $p$ and $q$; (2) the similarity of intrinsic geometric structure between $p$ and $q$. Then the similarity between $p$ and $q$, $M(p; q)$, can be computed as

$$M(p; q) = M_v(p; q) + \lambda M_g(p; q) \tag{1}$$

where $M_v(p; q)$ is the relevancy between $p$ and $q$ in visual words, $M_g(p; q)$ is the similarity between $p$ and $q$ in intrinsic geometric structure, $\lambda$ is the weight.

Assuming that $p$ is a bundled feature in clothing image $C_a$, then its matched bundled feature in clothing
image $C_b$ is $m(p) = \arg\max_{b \in C_a} M(p;b)$

So we can get a feature mapping $f_{A \rightarrow B}(p)$ from clothing image $C_A$ to clothing image $C_B$:

$$f_{A \rightarrow B}(p) = \begin{cases} 0, & \text{if } \max_{b \in C_a} M(p;b) = 0 \\ m(p), & \text{if } \max_{b \in C_a} M(p;b) > 0 \end{cases} \quad (2)$$

How to calculate $M_a(p;q)$ and $M_p(p;q)$ becomes the key step to construct a reasonable feature mapping $f_{A \rightarrow B}(p)$.

### 3.1 Computing $M_m$ based on LWF

Assuming that $p$ and $q$ are the two bundled features to be matched, when they relate to a number of common visual words, they are visually similar in certain degree. Obviously, when the number of common visual words is larger, then similarities between $p$ and $q$ would be higher.

In Zhong et al.(2009) and Zhipeng et al.(2010) papers, $M_m(p;q)$ is simply computed as the number of common visual words, which however cannot distinguish the situations that a common visual word relative to different number of SIFT features in $p$ and $q$, as mentioned in section 1. In our approach, to each common visual word $w_i$, we calculate its frequencies appeared in $p$ and $q$, $f_{w_i}$, respectively, which form a frequency pair, $F_{w_i} = \{f_{w_i}, f_{w_i}\}$. Then $M_m(p;q)$ is defined as

$$M_m(p;q) = \sum_{i=1}^{L} \left[ \frac{f_{w_i} + f_{w_i}}{2} - \sqrt{f_{w_i} f_{w_i}} \right] \quad (3)$$

where $L$ is the size of $\{w_i\}$.

We use Figure 3 to explain the effectiveness of equation (3). $M'^{I}_m$ is the value of $M_m(p;q)$ in Zhong et al.(2009) and Zhipeng et al.(2010), $M''_m$ is the value of equation (3). In the figure, we use triangle, round cake, diamond and squares to represent four visual words respectively. Three common words, i.e. triangle, round cake and squares are appeared both in $p$ and $q$. In such a situation, Zhong et al.(2009) and Zhipeng et al.(2010) return the same value of $M_m(p;q)$ for Figure 3(a) and Figure 3(b), which however is obviously not exact. While equation (3) can distinguish the difference of $M_m(p;q)$ in Figure 3(a) and Figure 3(b) effectively.
Figure 3: Computing $M_m$.

(a) $M'_m = 3, M''_m = 2e^{-2/3} + 2.5e^{-1/3} + 1.5e^{-1/2}$
(b) $M'_m = 3, M''_m = 1 + 2e^{-2/3} + 1.5e^{-1/2}$

3.2 Computing $M_g$ based on SDM

Geometric verification is an important step for improving image retrieval precision. Obviously, when we rotate or scale the clothing image, the information of clothing does not change, which requires the bundled features should be rotation and scale invariant.

Figure 4: The presentation of matched point pairs.

We propose a SIFTs distance matrix (SDM) to represent the inner geometric structure among SIFT features in
a bundled feature. Supposing that \( \{p; q\} = \{\{s_i, s_j\}\} \) is the set of matched SIFT feature pairs in bundled features \( p \) and \( q \), as shown in Figure 4. Then \( \{s_i\} \) and \( \{s_j\} \) in \( \{\{s_i, s_j\}\}\) can be written in two vectors, \( P_{n \times 1} \) and \( Q_{n \times 1} \), respectively, and \( \{p; q\} \) can be represented as \( [P_{n \times 1}, Q_{n \times 1}] \).

**Definition 1.** Let \( V_{n \times 1} = [v_1, v_2, \ldots, v_n]^T, v_k \in R^2 \), then

\[
V_{n \times 1} \otimes V_{n \times 1}^T = [v_1, v_2, \ldots, v_n]^T \otimes [v_1, v_2, \ldots, v_n] = \begin{bmatrix}
    d_{11} & d_{12} & \cdots & d_{1n} \\
    d_{21} & d_{22} & \cdots & d_{2n} \\
    \vdots & \vdots & \ddots & \vdots \\
    d_{n1} & d_{n2} & \cdots & d_{nn}
\end{bmatrix}
\]

where \( d_{ij} \) is the Euclidean distance of \( v_i \) and \( v_j \), thus the diagonal elements in the matrix are zero.

With definition 1, the inner geometric structure of \( p \) and \( q \) can be represented as SIFTs distance matrices:

\[
W_p = P_{n \times 1} \otimes P_{n \times 1}^T, W_q = Q_{n \times 1} \otimes Q_{n \times 1}^T
\]

**Definition 2.** We suppose that \( A_{n \times n}, B_{n \times n} \) are two \( n \times n \) matrices, then \( A \oplus B = \langle A, B \rangle = \frac{1}{\|A\| \|B\|} \), where \( \langle A, B \rangle = tr(B^T A) \), \( \|A\| = (A, A)^{1/2} \).

Accordingly we define \( M_g(p; q) \) as:

\[
M_g(p; q) = W_p \oplus W_q \quad (4)
\]

When we rotate an image, its inner geometric derived from definition 1 would not be changed, so the value of \( M_g(p; q) \) would not be changed in rotating its relative images, i.e. \( M_g(p; q) \) is rotation invariant. We also have proven that \( M_g(p; q) \) has the merit of scale invariance, as follows.

**Proof** We assume that the original image \( I(x, y) \) is given a scale parameter \( \alpha \), then we can get a scale transform image \( \hat{I}(x, y) = I(\alpha x, y) \) with the given scale transform matrix \( M = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix} \). Consequently, the coordinate of SIFT points in image \( I(x, y) \) and \( \hat{I}(x, y) \) satisfy \( x = \alpha x', y = \alpha y' \), we obtain a distance.
matrix $W$ of bundled feature in image $(x,y)$ and a corresponding distance matrix $W$ in image $(x,y)$, $W = \alpha W'$, then

$$\langle W'_p, W'_q \rangle = tr(W'_p W'_q) = \frac{1}{\alpha^2} tr(W'_p W'_q) = \frac{1}{\alpha^2} \langle W'_p, W'_q \rangle$$

$$W'_p \oplus W'_q = \left( \frac{1}{\alpha} (W'_p, W'_q) \right)^{1/2} = \left( \frac{W'_p, W'_q}{\alpha^2} \right)^{1/2} = \left( \frac{W'_p, W'_q}{\alpha^2} \right)^{1/2} = W_p \oplus W_q$$

We use Figure 5 to evaluate the effectiveness of equation (4). $M'_{g}$ is the value of $M_{g}(p,q)$ in Zhipeng et al. (2010) and $M''_{g}$ is the value of equation (4). In the figure, the areas of SIFTs-formed triangles in $p_1$ is equal to $q_1$, and the areas of SIFTs-formed triangles in $p_2$ is equal to $q_2$; however the triangle shapes in $p_2$ is different from the triangle shapes in $q_2$. If we follow Zhipeng et al.(2010), we have

$$M'_{g}(p_1; p_1) = M'_{g}(p_1; q_1), M'_{g}(p_2; p_2) = M'_{g}(p_2; q_2)$$

When we adopt our approach, we have

$$M''_{g}(p_1; p_1) > M''_{g}(p_1; q_1), M''_{g}(p_2; p_2) > M''_{g}(p_2; q_2)$$

which well explains the truth that the similarity between a bundled feature and itself should be higher than the similarity between two different bundle features. i.e. our approach is more robust than theirs in distinguishing similar SIFTs-formed triangle areas but different shapes.

We use Figure 5 to evaluate the effectiveness of equation (4). $M''_{g}$ is the value of equation (4). In the figure, the areas of SIFTs-formed triangles in $p_1$ is equal to $q_1$, and the areas of SIFTs-formed triangles in $p_2$ is equal to $q_2$; however the triangle shapes in $p_2$ is different from the triangle shapes in $q_2$. If we follow Zhipeng et al.(2010), we have

$$M'_{g}(p_1; p_1) = M'_{g}(p_1; q_1), M'_{g}(p_2; p_2) = M'_{g}(p_2; q_2)$$

When we adopt our approach, we have

$$M''_{g}(p_1; p_1) > M''_{g}(p_1; q_1), M''_{g}(p_2; p_2) > M''_{g}(p_2; q_2)$$

which well explains the truth that the similarity between a bundled feature and itself should be higher than the similarity between two different bundle features. i.e. our approach is more robust than theirs in distinguishing similar SIFTs-formed triangle areas but different shapes.

4. Clothing image similarity computation

For a query clothing image $C_q$ and its bundled features $C_q = \{b_{a1}, b_{a2}, \cdots, b_{am}\}$, the matched bundled feature
pairs between $C_A$ and a random clothing image $C_B$ in database with its bundled features $C_B = \{b_{i1}, b_{i2}, \cdots, b_{i_m}\}$ can be established by using equation (2). With the matched bundled feature pairs, the similarity between $C_A$ and $C_B$ is computed as the weighted sum of similarity between each bundled feature pair.

$$\text{Sim}(C_A, C_B) = \frac{\sum_{i=1}^{m} \eta_i \cdot M(b_{i}, f_{A,B}(b_{i}))}{\log(T_B)}$$ (5)

where $m$ is the number of bundled features in $C_A$, weight $\eta_i$ indicates the importance of the $i$-th bundled feature in $C_A$, $M(b_{i}, f_{A,B}(b_{i}))$ measures the similarity between $b_{i}$ and its corresponding bundled feature in $C_B$, $T_B$ is the number of SIFT features in $C_B$ which indirectly reflects the percentage of the matched bundled features against the total bundled features in $C_B$.

Computing $\eta_i$

Similar to text retrieval, the higher frequency of a word appeared in a text, the higher relevance between the word and the text content. Therefore, we can use the local word frequency to measure the importance of each visual word. The precision of the quantification indicates the relevant degree between each SIFT feature and its corresponding visual word. Thus we use the visual word importance and the SIFT feature quantification precision to define the importance of each SIFT feature.

For a query clothing image with its bundled features $C_A = \{b_{i1}, b_{i2}, \cdots, b_{i_m}\}$, its SIFT features $S = \{s_1, s_2, \cdots, s_T\}, s_i \in R^Z$, ($Z$ is the dimension of SIFT descriptor), and the vocabulary $W = \{w_1, w_2, \cdots, w_N\}, w_i \in R^Z$, supposing that the quantification of $S$ is $W = \{w_1, w_2, \cdots, w_L\}, 0 < L \leq N$, and the LWF of each $w_i$ is $f_i$, then the importance of $w_i$ can be measured as

$$\delta_k = \frac{f_k \cdot v_{k,\text{tfidf}}}{\sum_{i=1}^{L} f_i}$$

Where $v_{k,\text{tfidf}}$ is standard tf-idf weight[J. Sivic et al.(2003)]. For $s_i \in S$, supposing that its corresponding visual word is $w_j$, then the importance of $s_i$ is computed as

$$\lambda_k = \begin{cases} 0, & \text{if } s_i \cdot w_j \leq 0 \\ (s_i \cdot w_j) \cdot \delta_j, & \text{if } s_i \cdot w_j > 0 \end{cases}$$
Thus for a bundled feature \( b_i \) with its relative SIFTS \( \{s_i, s_{i+1}, \ldots, s_n\} \), its weight \( \eta_i \) in equation (5) can be computed as

\[
\eta_i = \sum_{j=1}^{n} \lambda_j
\]

\( \eta_i \) is further normalized as \( \eta_i = \frac{\eta_i}{\sum_{j=1}^{m} \eta_j} \) where \( m \) is the number of bundled features on the clothing image.

5. Experimental results

For evaluating our approach, we have taken pictures for 10 apparel under a variety of states, as shown the bottom row of images in Figure 6. Each apparel has 100 images. We have also downloaded 9,000 clothing images from internet(fuzhuang.taobao.com). Therefore, there are totally 10,000 images in the database. We clustered all the SIFT features on these images to form a vocabulary with about 10,000 visual words with hierarchical k-means clustering and use a soft quantization scheme (Philin et al., 2008) to quantize SIFT features into visual words. Then we use inverted-file index (W. Bruce et al., 2009) for indexing and retrieval.

![Mean Average Precision](image)

**Figure 6: MAP of different methods**

In experiments, we build four databases with different capacities (10,000, 20,000, 40,000 and 80,000) and all the SIFT features of the images in the databases are quantified into the 10,000 visual words mentioned before. Then we select 10 images of each clothing collection as the query image to retrieve the image databases by respectively using Bundled+Area (Zhipeng et al. 2010), our approach with NoWeight (i.e. \( \eta_i = 1 \) in equation (5))+Bundled+LWF+SDM and Weight+Bundled+LWF+SDM. Then for each database, we can get its Mean Average Precision (MAP) (W. Bruce et al., 2009), as shown in Figure 6. In experiments, the weight \( \lambda \) in equation (1) and \( \lambda \) in Zhipeng et al. (2010) are both set to 2. From the experimental results, it’s easy to get that the similarity measurement of NoWeight+Bundled+LWF+SDM is better than the approach in Zhipeng et al. (2010), and the Weight+Bundled+LWF+SDM similarity measurement is the best. Figure 7 gives a retrieval...
example and its Precision-Recall graph (W. Bruce et al., 2009) which indicates that our approach is effective in another criterion. The left is the query image on which the apparel region is dragged by the user, and the middle is the precision-recall, then the right is the search result. Figure 8 shows more example results using our approach with Weight+Bundled+LWF+SDM.

Figure 7: An example query and its precision-recall.

Figure 8: Example results. Sorting the clothing images in the database that have high visual similarity with the query clothing image (the left up one).

6. Conclusions

In this paper, we proposed a novel clothing image similarity computation method based on local features. We select the bundled features together with geometric constraints as clothing image features which have the merits of scale and rotation invariant. The bundled feature is motivated by SIFT features and maximally stable extremal region. SIFT features are quantified into a set of visual words for better efficiency. We compute the similarity between two bundled features from two aspects. One is by calculating the local word frequencies and their difference, which overcomes the problem of low discrimination between two regions with the same common visual words but different word frequencies; the other is by comparing the difference between the SIFTs distance matrices, which overcomes the problem of low discrimination between two regions that have similar SIFTs-formed triangle areas but different shapes. The whole clothing image similarity is computed by the weighted sum of similarities of the bundled features. The technique developed in this paper could be potentially used in online clothing retrieval.

We realize that some improvement could be made on current approach, e.g. we believe that setting the geometric constraints between bundled features would make the algorithm more robust. We plan to make such improvements in the future work.
Reference


W. Bruce Croft, Donald Metzler and Trevor Strohman.(2009) Search Engines: Information Retrieval in Practice. Addison-Wesley, Boston, USA.


