Efficient Clothing Retrieval with Semantic-Preserving Visual Phrases

Jianlong Fu[1], Jinqiao Wang[1], Zechao Li[1], Min Xu[2], and Hanqing Lu[1]

[1] Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2] Centre for Innovation in IT services and Applications
University of Technology, Sydney, Australia
{jlfu,jqwang,zcli,luhq}@nlpr.ia.ac.cn, min.xu25@gmail.com

Abstract. In this paper, we address the problem of large scale cross-scenario clothing retrieval with semantic-preserving visual phrases (SPVP). Since the human parts are important cues for clothing detection and segmentation, we firstly detect human parts as the semantic context, and refine the regions of human parts with sparse background reconstruction. Then, the semantic parts are encoded into the vocabulary tree under the bag-of-visual-word (BOW) framework, and the contextual constraint of visual words among different human parts is exploited through the SPVP. Moreover, the SPVP is integrated into the inverted index structure for accelerating the retrieval process. Experiments and comparisons on our clothing dataset indicate that the SPVP significantly enhances the discriminative power of local features with a slight increase of memory usage or runtime consumption compared to the BOW model. Therefore, the approach is superior to both the state-of-the-art approach and two clothing search engines.

1 Introduction

Large scale clothing retrieval has been attracted more attention in recent years. The keywords based search using "blue, T-shirt, short sleeve" has been provided in some online shopping websites, such as Amazon.com and eBay.com. Other emerging online shopping websites such as pixcoo.com and taotaosou.com have provided the content based similar clothing search by user interaction, e.g., drawing the region of clothing and selecting a clothing category.

In general, the clothing images on the shopping sites have clean backgrounds, while the images captured by digital cameras or mobile phones are with clutter backgrounds. So there are big discrepancies between the two scenarios. Additionally, the various human poses also increase the challenges.

The bag-of-words (BOW) model introduced in [1] is one of the most popular image representations for image retrieval.However, this representation may cause a great quantity of false correspondences due to the lack of semantic information and context constraint. For clothing retrieval, we can effectively utilize the semantic information of human parts and their relationships. To address the clothing retrieval problem in large scale, we propose a semantic-preserving visual phrases (SPVP) representation, which encodes the semantic partition and

K.M. Lee et al. (Eds.): ACCV 2012, Part II, LNCS 7725, pp. 420–431, 2013.

explores the word correlation between different semantic parts. Moreover, we show that the SPVP can be integrated into the inverted index structure that is efficient to be applied to the large scale dataset. Our approach exploits the content based image retrieval (CBIR) techniques and does not need the user interaction, therefore it can provide a more convenient access to online shopping and bring more commercial benefits.

The main contributions can be summarized as follows:

1. To eliminate the background noise and remove false detections, we refine each human part by sparse reconstruction with surrounding background regions outside human parts in the same image.
2. We propose a semantic-preserving visual phrases (SPVP) to model the word co-occurrence across different semantic parts, which could effectively combine the semantic constraint and spatial correlation.
3. Through embedding the semantic parts into vocabulary tree, the SPVP could be integrated into the inverted index structure for accelerating the retrieval process. The performance and speed outperform the state-of-the-art approach and two clothing search engines.

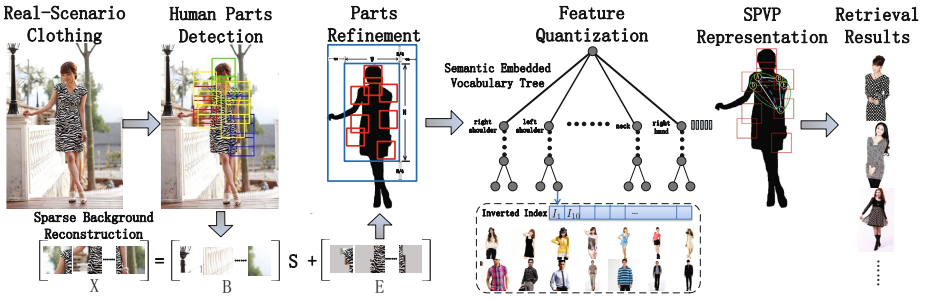

Fig. 1. Overall framework of clothing retrieval with semantic-preserving visual phrases. The squares in "Human Parts Detection" module are the human parts and we define the "surrounding background" as the area between the two blue rectangles in "Parts Refinement" module. In "semantic embedded vocabulary tree" module, the first layer is the semantic layer and the others are the feature layers.

2 Related Work

Clothing segmentation, modeling and recognition have been a long term research problem. One representative work for clothing modeling is an And-Or graph representation [2] for clothing configurations. Hasan et al. [3] and Wang et al. [4] proposed to utilize the priors to segment clothing. Yu et al. [5] proposed to take advantages of face detection, tracking and an efficient clothing segmentation method to recognise clothing in surveillance videos. Zhang et al. [6] leveraged both low-level features and high-level attributes for clothing search. Huang et.al [7] designed the color features, shape features and statistical features for suit

detection. Recently, Liu et al. [8] proposed an unsupervised transfer learning method to model the cross-domain learning problem in clothing retrieval.

The other related works are the BOW model and its extension. Most state-of-the-art large scale retrieval approaches are based on the BOW model with inverted files to implement fast searching. Due to the ignoring of the spatial information in traditional BOW model, numerous works have been proposed to successfully improve the retrieval performance. Some post-processing methods like RANSAC are applied to the top images in the initial ranking, but these methods introduce more expensive computational costs. The alternative to this is adopting the weak geometry constraints such as the visual synset [9] and spatial bag of features [10]. These methods based on weak spatial information are usually inadequately to model the spatial layout. Another approach is to use visual phrases based on the co-occurrences in the local regions or entire images. Bundling features [11] and phrases [12] capture local information without long range interactions. To overcome these issues, Zhang et al. [13] proposed an interesting and promising method to model the spatial layout of visual words. However, their representation only captures the translation invariance but ignores the scale or rotation invariance. Hence, how to efficiently encode the spatial or even semantic information in large scale image retrieval is still a remaining challenge.

3 Generation of Semantic Parts

3.1 Human Parts Detection

Detection based human body parsing technique [14] is proven to be helpful in human parts location and matching. Therefore, we use the flexible mixture parts model [14] to train both a upper body detector and a lower body detector. For the sake of efficiency, we detect 18 upper body parts and 8 lower body parts. The example of upper body detection is shown in Fig. 1. After the human parts alignment, two corresponding human parts like the left shoulder which are located in two images can be linked up and share similar feature representation which is important for the following steps.

3.2 Semantic Parts Refinement

For each human part, there exist some false detections and background regions. Therefore, further feature refinement process to eliminate the background influence is necessary. Since it is difficult to guarantee good segmentation results and afford the computational expenses for realtime applications, we solve the problem by a sparse background reconstruction with effectiveness and efficiency. We refine human parts by projecting them into a linear subspace which is constructed by background in the same feature space. We assume that the background is consistent in a small range. Therefore, the background both in human parts and the regions outside the parts in the same image tend to share the identical feature descriptions. Then we can reconstruct the background area in each part using

surrounding background with low reconstruction error. In practice, we extend one quarter of the width and length of the bounding box of all the human parts in an image and define the "surrounding background" as the area between the two blue rectangles, as shown in Fig. 1. We extract features from surrounding background and regard them as the basis vector. Theoretically, since the background is consistent within a small range, the reconstruction error of background in human parts is very small and the error of foreground clothing is large. Additionally, the background is usually low-rank and has some repetitive structures, so we can guarantee the basis vector is adequate to reconstruct the background in parts for most situations. That is to say, the basis vectors are over complete.

With the above analysis, we denote $B = [b_1, b_2, \ldots, b_n] \in R^{d\times n}$ as features extracted from surrounding background and $X = [x_1, x_2, \ldots, x_m] \in R^{d\times m}$ as features from human parts, where each column of B and X is a feature vector. To implement the feature extraction from both the surrounding background and human parts, we partition each region into multi-scale grids, i.e., 8×8 and 4×4. Six kinds of features (details in the experiment) extracted from each grid are finally concatenated. We learn a sparse coefficient matrix $S \in R^{n\times m}$ to reconstruct X using B with reconstruction error $E \in R^{d\times m}$. The objective function is formulated as:

$$\min_S \frac{1}{2}||X - BS||_2^2 + \lambda||S||_1 \tag{1}$$

The l_1 norm of a matrix is the sum of the absolute values of its entries [15]. The fact that features can "cancel each other out" using subtraction is contrary to the intuitive notion of combining parts to form a whole [16,17]. Therefore, we constrain S to be non-negative. The above optimization problem has a closed form solution:

$$S = (B^T B)^{-1}(B^T X - \lambda) \tag{2}$$

Where the $B^T B$ is a $n\times n$ matrix. Since the n is not very large (in our experiment, n = 640), the calculation of $(B^T B)^{-1}$ is fast, and the optimization problem can be solved quickly. Then we use the following equation to get the reconstruction error:

$$E = X - BS \tag{3}$$

If the error E is near to zero, we regard the grid as background, otherwise we regard it as foreground. Some examples of the reconstruction map will be given in the experiment section.

4 Semantic Embedded Vocabulary Tree

For large scale image retrieval, the employment of local invariant features [18] and BOW model [1] is a successful framework. The hierarchical vocabulary tree [19,20] is widely used for feature quantization. However, the visual words are only quantized based on low-level local features, without considering any semantic information in the vocabulary tree. Therefore, we embed the semantic parts into the vocabulary tree structure.

We propose to incorporate the semantic parts by adding a semantic layer to the traditional vocabulary tree structure. The quantization process is a two-step scheme to quantize a feature into the semantic embedded vocabulary tree. We denote a tree T as a concatenation of the semantic layer and the feature layer, as shown in Fig.2. The depth of the semantic layer is 1 and the number of branches is determined by the number of semantic parts. In the semantic layer, the visual features are mapped into the semantic nodes according to the parts partition. For the feature layers, the hierarchical quantization technique is employed to map the feature space. Each subtree of a semantic node is a traditional hierarchical vocabulary tree obtained by hierarchical K-means clustering of local features, with a branch factor K and depth L (in our experiment, the depth of the hierarchical vocabulary tree is set to 6 with branch factor 10). In the l^{th} layer, node n^{l,h_l} is a visual word representing the clustering center, in which h_l denotes the node index in this layer. An image is represented by a bag of visual features $\{x_i\}$. Each feature x_i is quantized to the corresponding visual words by traversing the root node to leaf node of T and can be eventually represented by the set of nodes, *i.e.*, $T(x_i) = \{n_i^{l,h_l}\}_{l=1}^{L_i}$, in which n_i^{1,h_1} denotes the index in the semantic layer and n_i^{l,h_l} $(l \geqslant 2)$ denotes the index in the feature layers. In this way, the semantic parts are effectively embedded into the feature space, which could be further used to visual search.

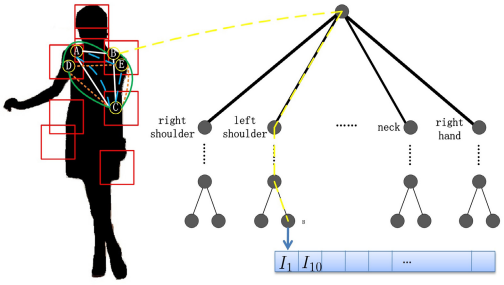

Fig. 2. Semantic embedded vocabulary tree and the SPVP representation

5 Clothing Retrieval with SPVP

5.1 Semantic-Preserving Visual Phrase

The visual words in different human parts are related to each other in terms of representing the characteristic of texture distribution, which results from different design styles of clothing. Matching only one part can not provide users with satisfied results. Therefore, we select M words from M different human parts respectively to form the semantic-preserving visual phrase (SPVP), where M is the length. This could provide a contextual constraint for robust clothing retrieval.

An image will be represented as a vector defined with the SPVP. Similar to the BOW model, the vector $V^M(I)$ of an image is defined as the histogram of SPVP of length M. M is the number of correlated visual words, each dimension

of $V^M(I)$ is a phrase with its frequency. Each phrase is a possible combination of M words from different semantic parts. More specifically, if there are M words separately located in M different semantic parts, it corresponds to a SPVP of length M. As shown in Fig. 2, if we calculate the SPVP of length 3, we can choose 3 words in 3 different semantic parts and find that the ABC, AEC, DBC, DEC are the four SPVP of length 3. The SPVP representation of larger length requires that the corresponding words occur in more different semantic parts rather than within one or two parts.

5.2 Searching with SPVP

In the following, we describe the voting procedure for a query q to obtain the similarity scores with database images through SPVP:

1. Initialize the score of each database image to zero in all semantic parts.
2. Traverse all parts and obtain the number of matched words for each part separately. If a word j appears both in the p^{th} part of the query q and the p^{th} part of a database image I, we increment the score of image I in the p^{th} part.
$$S_{I,p} = S_{I,p} + 1 \tag{4}$$
3. Denote P as the number of the semantic parts, P' as the subset of P in which $S_{I,p} > 0$. To select M candidate parts which contain one matched word at least, we combine M different semantic parts from the set P', which is denoted by the set $\Phi = \{\phi_1, \phi_2, ..., \phi_{C_{P'}^M}\}$. For ϕ_k, $k = \{1, 2..., C_{P'}^M\}$, it corresponds to a kind of part combination. To calculate the term frequency of a SPVP of length M, we can select one word from each part respectively to form a SPVP. By traversing all possible combinations, we can get the total phrase set, which is denoted by $\Theta_k = \{\theta_{k1}, ..., \theta_{ki}, ..., \theta_{kz}\}$, z is the number of unique phrases. Noted that one word can occur several times in each part, so one phrase can also occur several times in M parts, which denoted as $tf(\theta_{ki})$.
4. To calculate the idf term, we define the idf of a phrase θ_{ki} as the sum of the idf term of the words belonging to it. And the value for a word w_j is calculated as $log(N/N_j)$, where N_j is the number of images that contain w_j and N is the total number of images.
$$idf(\theta_{ki}) = \sum_{w_j \in \theta_{ki}} idf(w_j) \tag{5}$$
5. The similarity score of the query q and image I is defined as:
$$Score(q, I) = \sum_{k=1}^{C_{P'}^M} \sum_{i:\theta_{ki} \in \Theta_k} tf(\theta_{ki}) \times idf(\theta_{ki}) \tag{6}$$

5.3 Index and Storage

To reduce the storage and accelerate the retrieval speed, an inverted index structure is usually adopted for large scale indexing and searching. The traditional BOW representation with inverted index only calculates the word frequency but ignores the semantic information in the original image space [21], which often leads to a large number of false matchings and bias the true similarity scores.

We incorporate the semantic parts into the inverted index structure for large scale indexing and retrieval. Fig. 3 shows the structure of our index. In addition to the image ID, for each occurrence of a visual word, we use P semantic bins to respectively record the amount of features which belong to the same word but located in different semantic parts. In practice, for each entry of a visual word, we use 19 bits for the image ID and 7 bits for each semantic bin. More details will be given in the experiment to compare the memory usage between the traditional BOW and SPVP.

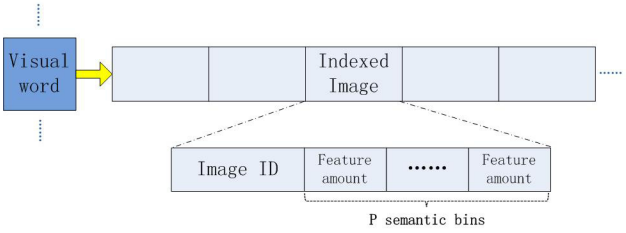

Fig. 3. Improved inverted index structure

6 Experiment

6.1 Dataset and Experimental Setting

There has been already several clothing datasets, proposed by Yang et al. [5], Bourde et al. [22], T.chen et al. [23] and Liu et al. [8]. However, most of them are not suitable to our task because of the low resolution and few attributes. [8] is fine, but it is not available to the public. Therefore, we collect two datasets similar to [8], namely Online Shopping (OS) dataset and Daily Photo (DP) dataset.

The OS dataset is used as index, which is collected by crawling images with clean background from Amazon.com, eBay.com, vancl.com, taobao.com by typing clothing categories such as "male, T-shirt", "female, trousers" as query keywords. The DP dataset is presented to the system as queries, which is collected from google image and Flickr.com using queries such as "real scenario, male, T-shirt", "real scenario, female, trousers", etc. Totally, we collect 110 thousand images in OS dataset and 2 thousand images in DP dataset. More specifically, the number of upper body is 70 thousand and the lower body is 40 thousand in OS. Additionally, we collect 1 thousand upper body images and 1 thousand lower body images in DP dataset for testing. The category of DP is the same as the OS dataset.

We annotate a set of clothing-specific attributes as [8]. The attributes are summarized into three classes including global, upper body and lower body attributes. The global attributes can be divided into color and pattern, the upper body attributes can be divided into sleeve, collar and front, the lower body attributes can be divided into shape, length, curling and drape.

We measure the clothing retrieval performance by the evaluation criterion of [8] and [24]. Let $Rel(i)$ be the ground truth relevance between the query image q and i^{th} ranked image in OS dataset. We can evaluate the top k retrieved datum by a precision.

$$Precision@k = \frac{\sum_i^k Rel(i)}{N} \tag{7}$$

Where N is a normalization constant to ensure the correct ranking results in a precision score of 1. If we evaluate on multiple attributes of a query image, the $Rel(i)$ will have multiple levels of relevance values.

6.2 Within-Scenario versus Cross-Scenario

To testify the influence of background discrepancy in cross-scenario, we randomly select 200 queries from the DP dataset. To simulate the clothing image with clean background, we manually extract the foreground clothing as the within-scenario query. The original image with clutter background is regarded as the cross-scenario query.

We also test the part based approach with sparse background reconstruction. We use 18 upper body parts and 8 lower body parts for retrieval and set the length of the SPVP to 2. We will demonstrate that the length 2 is the best one among the lengths from 1 to 5 in the next section. To enhance the matching robustness, we extract 6 kinds of features from multi-scale grids in parts, i.e., 8×8 and 4×4. The features cover HOG, LBP, SIFT, Gabor texture, skin descriptor and dominant color (DC) descriptor in HSV color space. We adopt the decision fusion strategy for several different types of feature. For DC, we quantize color into 28 bins and set the quantized colors of bins with $M_i/M > \gamma$ to one, where M_i is the feature number of the i^{th} bin, M is the sum of all the bins and γ is the predefined threshold (e.g., $\gamma = 0.4$). DC is designed to robust against the disturbance of small noises. Fig. 4 shows the retrieval performances of within-scenario, cross-scenario with and without sparse background reconstruction. We can see that the performances drop from within-scenario to cross-scenario at about 18% and 21% for the upper body and lower body clothing respectively. It can also be seen that the number decreases to 4% and 10% with the help of the sparse background reconstruction. The experiment shows the large discrepancies between the within-scenario and the cross-scenario, as well as the retrieval performance boosted by the sparse background reconstruction.

6.3 Performances with Different Phrase Length

We examine the effect of the phrase length in our approach without the sparse background reconstruction. From Fig. 5, the experiment shows that length $M =$

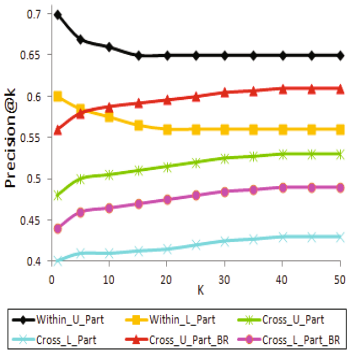

Fig. 4. The performance of within-scenario, cross-scenario with and without background reconstruction(BR). The letter U and L stands for the upper and lower body respectively. "Part" means the approaches are based on human parts alignment.

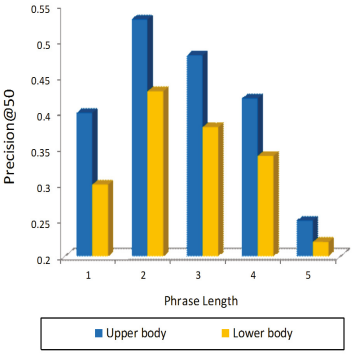

Fig. 5. Precision@50 with SPVP of different length. Length 2 is the optimal length from 1 to 5 for both upper body and lower body images.

2 is the best from 1 to 5. The SPVP with length 1 is degraded to the BOW model. It can be seen that the SPVP is better than the BOW model from length 1 to length 4, for the words correlation is considered. We can also observe that a longer phrase stands for a rigorous global matching and it will lead to a lower recall in the retrieval process. As shown in Fig. 5, the length $M = 5$ means that if there are no 5 matched words in 5 different parts respectively, the approach sets the similarity score of two images to zero.

6.4 Comparison with State-of-the-Art

To demonstrate the effectiveness and efficiency of our approach, we compare the SPVP based approach with several clothing search engines and the "street to shop" approach [8]. In general, in some online shopping websites, users are asked to draw the region where the clothing are located. We firstly set three strong baselines by manually selecting the bounding box of the target clothing in real scenario image. We extract six kinds of features from the bounding box area as a whole. We also compare with image search engine "images.google.com" and clothing search engine "pixcoo.com". For the three baselines, we use the region of clothing as query without the part alignment and refinement.

The experimental results are given in Fig. 6. We can see that our approach and the "street-to-shop" approach, are discriminative than the three global appearance based methods. There are 63% and 30% *Precision*@50 improvements at most for the upper body images and the lower body images. Additionally, our approach performs better for large scale clothing retrieval than [8]. Relying on the sparse background reconstruction, our SPVP based approach with length 1, which is degraded to BOW model, achieves better results than the

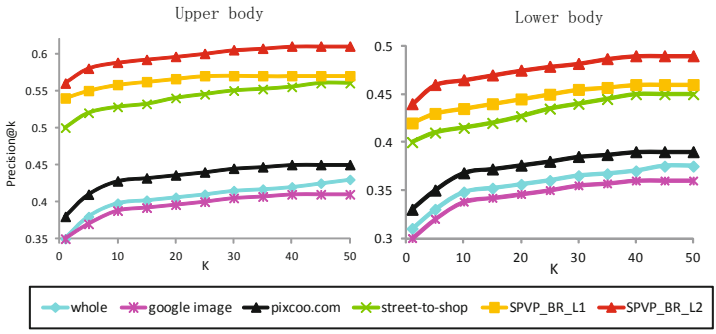

Fig. 6. Comparison results of our approach with state-of-the-art approaches. We compare the baseline using global features (whole), two web search engines, the "street-to-shop" and our approach with length one (SPVP_BR_L1) and length two (SPVP_BR_L2).

"street-to-shop" approach [8]. It shows that the sparse background reconstruction is more effective to the query image. The fact verifies our assumption that the background can share the similar property in a small range. Fig. 7 gives some examples of sparse background reconstruction with our approach. For normalization, the error of each grid is divided by the max error in one part. It can be seen that the grids with non-zero error represent the foreground clothing accurately. Fig. 8 shows some exemplar retrieval results. It can be seen that the SPVP based approach provides a promising result, since our approach guarantees the similarity of the most concerned attributes when people searching, such as the color, pattern, the length of sleeves and trousers.

Fig. 7. The examples of sparse background reconstruction. Our aim is to subtract the background in part or remove the false human part detection. The area within the red line indicates the grids with non-zero error(foreground).

Additionally, to measure the computational cost, we use several evaluation criterions to compare the BOW, SPVP and the "street-to-shop" approach. Our experiment runs on a PC with Intel Core Duo CPU with $2.53GHz$ and $3GB$

memory. Table. 1 shows the memory usage as well as the running time of the three approaches. Noted that the search time of SPVP covers two stages, the sparse background reconstruction and the SPVP voting. The proposed approach achieves a significant improvement in retrieval accuracy with a slight increase of time consumption and memory usage. On the contrary, the "street-to-shop" approach takes more than 3 seconds per query. The increased time consumption is mainly derived from the solving of the within-scenario spare representation coefficient and the nearest neighbor search.

Table 1. Comparison in the memory usage and average runtime

Approach	Storage	Memory usage	Runtime	
			Detection	Search
BOW	Inverted Index	198 MB	0s	0.20s
SPVP	Inverted Index	462 MB	0.3s	0.37s
Street-to-shop	Auxiliary Set	1.35 GB	0.3s	3.00s

7 Conclusion

We propose an approach that encodes semantic context into the BOW model to solve the cross-scenario clothing retrieval problem in large scale. Also, to further refine semantic parts, a fast solution of sparse background reconstruction is proposed with effectiveness and efficiency. The experiments show that our approach outperforms the state-of-the-art approach as well as two clothing search engines on the performance of retrieval accuracy, memory usage and runtime.

Fig. 8. Some retrieval exemplars of real scenario clothing images. The retrieved correct and wrong clothing attributes are marked with green stars and red crosses respectively.

Acknowledgement. This work was supported by the National Natural Science Foundation of China 973 Program (Project No. 2010CB327905) and the National Natural Science Foundation of China (Grant No. 60905008, 60833006).

References

1. Sivic, J., Zisserman, A.: Video google: a text retrieval approach to object matching in videos. In: ICCV (2003)
2. Chen, H., Xu, Z., Liu, Z., Zhu, S.: Composite templates for cloth modeling and sketching. In: CVPR (2006)
3. Hasan, B., Hogg, D.: Segmentation using deformable spatial priors with application to clothing. In: BMVC (2010)
4. Wang, N., Ai, H.: Who blocks who: Simultaneous clothing segmentation for grouping images. In: ICCV (2011)
5. Yang, M., Yu, K.: Real-time clothing recognition in surveillance videos. In: ICIP (2011)
6. Wang, X., Zhang, T.: Clothes search in consumer photos via color matching and attribute learning. In: ACM Multimedia (2011)
7. Huang, L., Xia, T., Zhang, Y., Lin, S.: Finding Suits in Images of People. In: Schoeffmann, K., Merialdo, B., Hauptmann, A.G., Ngo, C.-W., Andreopoulos, Y., Breiteneder, C. (eds.) MMM 2012. LNCS, vol. 7131, pp. 485–494. Springer, Heidelberg (2012)
8. Liu, S., Song, Z., Liu, G., Xu, C., Lu, H., Yan, S.: Street-to-shop: Cross-scenario clothing retrieval via parts alignment and auxiliary set. In: CVPR (2012)
9. Zheng, Y., Zhao, M., Neo, S., Chua, T., Tian, Q.: Visual synset: Towards a higher-level visual representation. In: CVPR (2008)
10. Cao, Y., Wang, C., Li, Z., Zhang, L., Zhang, L.: Spatial-bag-of-features. In: CVPR (2010)
11. Wu, Z., Ke, Q., Isard, M., Sun, J.: Bundling features for large scale partial-duplicate web image search. In: CVPR (2009)
12. Yuan, J., Wu, Y., Yang, M.: Discovery of collocation patterns: from visual words to visual phrases. In: CVPR (2007)
13. Zhang, Y., Jia, Z., Chen, T.: Image retrieval with geometry-preserving visual phrases. In: CVPR (2011)
14. Yang, Y., Ramanan, D.: Articulated pose estimation with flexible mixtures-of-parts. In: CVPR (2011)
15. Zhang, Z., Liang, X., Ganesh, A., Ma, Y.: TILT: Transform Invariant Low-Rank Textures. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part III. LNCS, vol. 6494, pp. 314–328. Springer, Heidelberg (2011)
16. Hoyer, P.: Non-negative sparse coding. In: IEEE Workshop on Neural Networks for Signal Processing (2002)
17. Li, Z., Yang, Y., Liu, J., Zhou, X., Lu, H.: Unsupervised feature selection using nonnegative spectral analysis. In: AAAI (2012)
18. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV 60, 91–110 (2004)
19. Fu, J., Wang, J., Lu, H.: Effective logo retrieval with adaptive local feature selection. In: ACM Multimedia (2010)
20. Nisterand, D., Stewenius, H.: Scalable recognition with a vocabulary tree. In: CVPR (2006)
21. Fu, J., Wang, J., Zhang, Y., Lu, H.: Point-context descriptor based region search for logo recognition. In: ACM ICIMCS (2012)
22. Bourdev, L., Maji, S., Malik, J.: Describing people: a poselet-based approach to attribute classification. In: ICCV (2011)
23. Gallagher, A., Chen, T.: Clothing cosegmentation for recognizing people. In: CVPR (2008)
24. Siddiquie, B., Feris, R., Davis, L.: Image ranking and retrieval based on multi-attribute queries. In: CVPR (2011)