MOBILE VISUAL CLOTHING SEARCH

George A. Cushen and Mark S. Nixon

University of Southampton
{gc505, msn}@ecs.soton.ac.uk

ABSTRACT

We present a mobile visual clothing search system whereby a smart phone user can either choose a social networking photo or take a new photo of a person wearing clothing of interest and search for similar clothing in a retail database. From the query image, the person is detected, clothing is segmented, and clothing features are extracted and quantized. The information is sent from the phone client to a server, where the feature vector of the query image is used to retrieve similar clothing products from online databases. The phone's GPS location is used to re-rank results by retail store location. State of the art work focusses primarily on the recognition of a diverse range of clothing offline and pays little attention to practical applications. Evaluated on a challenging dataset, the system is relatively fast and achieves promising results.

***Index Terms*—** Clothes Search, Mobile Search, Image Retrieval

1. INTRODUCTION

Clothing was the fastest growing segment in US e-commerce last year, with it predicted to have grown by 20% to $40.9 billion from 2011 to 2012. It is also expected to have been the second biggest segment by revenue overall [1]. Thus, an efficient mobile application to automatically recognize clothing in photos of people and retrieve similar clothing items that are available for sale from retailers could transform the way we shop whilst giving retailers a great potential for commercial gain. Tightly connected to this, is the potential for an efficient clothing retrieval system to be employed for the purpose of highly targeted mobile advertising which learns what clothing a person may wish to purchase given their social networking photos.

The problem of efficient and practical mobile clothing search appears relatively unexplored in literature. Recently, the fields of clothing segmentation, recognition and parsing have started to gain much attention in literature. Gallagher and Chen designed a Graphcuts approach to segment clothing [2] to aid the person recognition application. Various priors have been proposed to segment clothing by Hasan and Hogg [3] and Wang and Ai [4]. Meanwhile Yang and Yu [5] proposed to integrate tracking and clothing segmentation to recognize clothing in surveillance videos. Although their method is fast, they capture their dataset in a controlled lab with a simple white background. The clothing retrieval problem has been less extensively studied. One scenario is presented in [6, 7]. State of the art work focusses primarily on clothes parsing and semantic classification [8, 9]. Although, Yamaguchi et al. achieve good performance, they only briefly demonstrate retrieval and their method is very computationally intensive.

Current mobile image retrieval systems include Google Goggles[1], Kooaba[1], and LookTel[1]. However, these systems are developed for image retrieval on general objects in a scene. When these systems are applied to clothes search, they can provide visually and categorically less relevant results than our method for retrieving products based on a dressed person and can have significantly longer response times than our method.

The main contributions of this paper are as follows: (1) we present a novel mobile client-server framework for automatic visual clothes searching; (2) we propose an extension of GrabCut for the purpose of clothing segmentation; (3) we propose a dominant colour descriptor for the efficient and compact representation of clothing; and (4) we have evaluated our approach on query images from a fashion social network dataset along with a clothing product dataset for results and shown promising retrieval results with a relatively fast response time. The contributions in this paper thus reside in a mobile system for automated clothes search with proven capability.

2. SYSTEM OVERVIEW

The pipeline for our mobile visual clothing search system to retrieve similar clothing products in nearby retail stores is shown in Figure 1. A smart phone user can either capture a photo of a person wearing clothing of interest or choose an existing photo, such as from a social network. The person is then detected in the image and our clothing segmentation method is performed to attempt to select only the clothing pixels for the next step of feature extraction. Note that we only consider searching upper body clothing since the images in our social

[1] `google.com/mobile/goggles`, `kooaba.com`, `looktel.com`

networking dataset indicate that many people just take upper body fashion photos. The segmented upper body clothing image is divided up into non-overlapping patches and dominant colour and HoG features are extracted. These sets of descriptors are quantized using vocabulary codebooks and concatenated to generate a histogram of visual words (HoVW). The HoVW defines the ultimate query which is compared to a database of HoVWs for clothing products from retailers. Finally, a similarity measure is applied to determine the most similar matches and these are re-ranked based on the GPS location of the user (obtained from the smart phone) and the location of the retailers, stored in the database.

It is not practical to store databases of a large number of clothing products from various retailers on the client. Thus, a client-server architecture is conceived for our mobile visual clothing search.

Our system is designed to be efficient with short response times and offer an interactive graphical user experience. The client communicates with the server using compressed feature information rather than a large query image. This allows for fast transmission on typical 3G mobile networks and has the additional benefit of distributing processing between client and server so that the server may handle more simultaneous search requests. Our contributions are described in the following sections.

3. CLOTHING SEGMENTATION

Clothing segmentation is a challenging field of research which can benefit numerous fields including human detection [10], recognition for re-identification [2], pose estimation [11], and image retrieval. Although the fast segmentation of a person's clothing in a photo appears effortless for a human to perform, it remains challenging for a machine due to the wide diversity and ever-changing nature of fashion, uncontrolled scene lighting, dynamic backgrounds, variation in human pose, and self and third-party occlusions. Additionally, difficult sub-problems such as face detection are usually involved to initialize the segmentation procedure.

The main objectives of this stage of the system are to automatically crop the image to the region of interest (the regions of the body below the head where clothes are typically located) and to eliminate both the background and skin from the image to constrain the regions where clothing features will be extracted from.

We propose converting the query image I_q to the more perceptually relevant YCrCb colour space and the corresponding illumination channel is normalized to help alleviate, to some extent, the non-uniform effects of uncontrolled lighting.

The Viola-Jones face detector is used to estimate the face size and location which are fed as parameters to initialise a human detector based on [12]. The detector yields a ROI for the full body pose excluding head, ROI_p, and a smaller upper body only region, ROI_u. We attempt to segment the person from the background within the bounding box ROI_p by using the popular GrabCut algorithm. GrabCut is based on graph cuts which have been shown to be reasonably efficient and to have good performance at segmenting humans [13].

We attempt to eliminate the skin from the segmented person by employing an efficient thresholding method. Chai and Ngan [14] reported that skin pixels on the face can be identified by the presence of a certain set of chrominance values in the YCrCb colour space and utilized for face detection purposes. Based on this work, we propose a thresholding method for the purpose of clothing segmentation that takes into account other skin pixels on the body. This can be more challenging as we find illumination on the face tends to be more uniform. Consider R_r and R_b as ranges of the respective Cr and Cb values that correspond to the colour of skin pixels. For a random sample of our social networking dataset, we found ranges of $R_r = [140\,165]$ and $R_b = [105\,135]$ to be optimal. In our experiments, these ranges prove to provide a good compromise between robustness against different types of skin colour and attempting to preserve clothing pixels of similar chrominance to the skin. Thus, we have the following equation:

$$\text{skin}(x,y) = \begin{cases} 1 & \text{if } \text{Cr}(x,y) \in R_r \cap \text{Cb}(x,y) \in R_b \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where x and y are pixels in ROI_p. Morphological opening is then performed on the binary mask $\text{skin}(x,y)$ to reduce noise.

Finally, the segmented full body clothing is cropped to the upper body region ROI_u and normalised in size. The area of segmented clothing is compared to the area of the ROI_u. If the percentage of clothing pixels is less than an empirically defined threshold τ_a, we perform the next stage (feature extraction) on the pre-processed image rather than the segmented image. This final step can increase overall robustness of the system to the special case where the clothing and either skin or background are of a very similar colour. The resulting upper body clothing image is denoted I_c.

4. CLOTHING FEATURE EXTRACTION

Colour is one of the most distinguishing visual features of clothing. We propose an efficient method to describe dominant colours in the segmented clothing image based on the MPEG-7 descriptor [15], and integrate this feature with the HoG texture/shape descriptor.

The upper body clothing image I_c is divided up into a regular grid of 5×5 cells (depicted in the third column of Figure 3). We denote each column of the grid as ROI_c^k where $k = 1 \ldots 5$ and we propose providing robustness to layered clothing (e.g. jacket and top) by computing the dominate colours for each column and concatenating.

A 3D histogram is computed on $I_c \in \text{ROI}_c^k$ in HSV colour

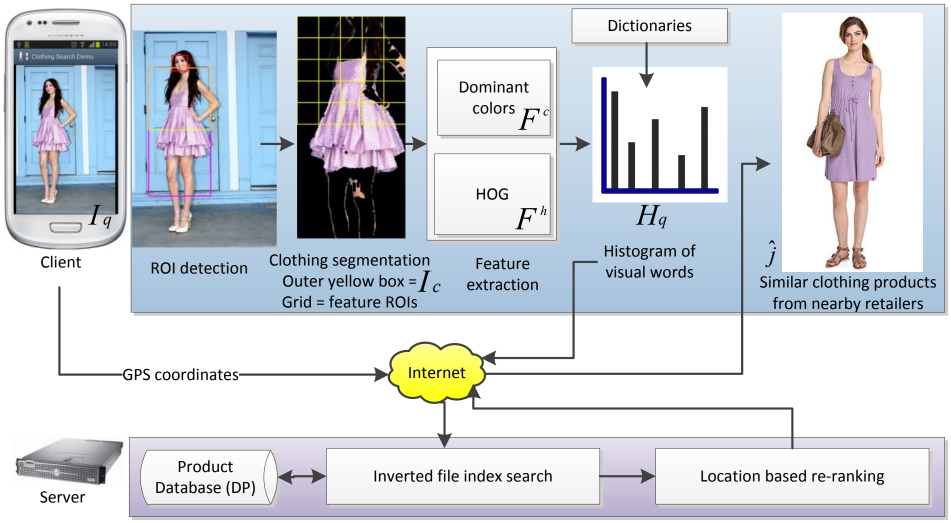


Fig. 1: Overview of our clothing retrieval pipeline.

space. For clothing, hue quantization requires the most attention. We find a quantization of the hue circle at 20° steps sufficiently separates the hues such that the red, green, blue, yellow, magenta and cyan are each represented with three subdivisions. Also, saturation and illumination are each quantized to three sub-divisions. Hence the colour is compactly represented with a vector of size $18 \times 3 \times 3 = 162$.

The quantized colour of each colour bin is selected as its centroid. If we let C_i represent the quantized colour for bin i, $X = (X^H, X^S, H^V)$ represent the pixel colour, and n_i be the number of pixels in bin i, we can calculate the mean of the bin's colour distribution as follows:

$$C_i = \bar{X}_i = \frac{1}{n_i} \sum_{j}^{n_i} X_{i,j} \quad , \quad 1 \leq i \leq 162 \tag{2}$$

Ideally, the dominant colours would be given by bins with the greatest percentage of image pixels. However, in practice, due to factors such as uncontrolled illumination, bins of similar quantized colours often exist per perceived clothing colour. Therefore, the mutual polar distance between adjacent bin centres is iteratively calculated and compared with a threshold, τ_d, and similar colour bins are merged using weighted average agglomerative clustering. Considering X_1 and X_2 in the adjacent bins, we let P_E represent the pixel percentage of the colour component E and perform the following equation for each colour component, substituting E for the H, S, and V components respectively:

$$X^E = X_1^E \left(\frac{P_1^E}{P_1^E + P_2^E} \right) + X_2^E \left(\frac{P_2^E}{P_1^E + P_2^E} \right) \tag{3}$$

Bins with a pixel percentage less than τ_p are considered insignificant colours and merged to their closest neighbour bin. Since each set of worn upper body clothing in our product dataset is humanly perceived to generally have less than 3 dominant colours per ROI$_c^k$, thresholds τ_d and τ_p are empirically defined to yield approximately this amount of dominant colours. For the purpose of our similarity stage, we convert the polar HSV colours to the Euclidean LAB space and the represent the dominant colours F_k^c as:

$$F_k^c = \{(C_1^L, C_1^A, C_1^B, P_1), \ldots, (C_n^L, C_n^A, C_n^B, P_n)\} \tag{4}$$

where (C_1^L, C_1^A, C_1^B) is a vector of LAB dominant colour, the corresponding percentage of that colour in the clothing is given by P_1 and $0 > n \leq 3$ is the number of dominant colours on the clothing. For our application, we generate $F^c = \{F_k^c\}$ (padding each F_k^c if necessary) to yield total dimensions of $4 \times 3 \times 5 = 60D$.

Texture/shape features based on histogram of oriented gradient (HoG) are computed in each cell on I_c globally quantized to its dominant colours. Gradient orientations are quantized to every 45°, thus there are 8 direction bins. The local histograms of the cells are then concatenated together to form the $8 \times 25 = 200D$ HoG feature F^h.

5. CLOTHING SIMILARITY

A Bag of Words (BoW) representation of the features is employed to increase robustness to noise, wrinkles, folding, and illumination. For a query image I_q, we perform the clothing segmentation and feature extraction steps and then the histogram of visual words, H_q, is generated as follows. For every image feature F^j, we locate its corresponding visual word w_n^j from every dictionary D_n. These visual words are accumulated into individual histograms H_n for each dictionary and the unified histogram is given by concatenating the individual histograms: $H_q = [H_1^T H_2^T]^T$.

Finally, an inverted index is employed, minimizing the L_1 distance between H_q and the codeword histogram H_j of the j^{th} clothing product in the product dataset to obtain the search result:

$$\hat{j} = \arg\min_{j} d_1(\mathbf{H_q}, \mathbf{H_j}) \tag{5}$$

where $d_1(\mathbf{H_q}, \mathbf{H_j}) = \|\mathbf{H_q} - \mathbf{H_j}\|_1 = \sum_{i=1}^{n} |H_q(i) - H_j(i)|$. This approach to searching is chosen as it allows for fast and efficient searching of large databases.

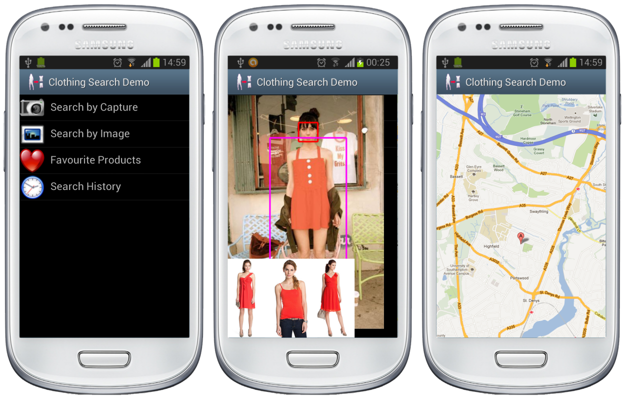
For training, the dominant colour and HoG features are extracted (as per our method for testing) from each image in the product database. A dictionary is built for each feature using Approximate K-Means. The codebook size is empirically set to 200 for F^c and 100 for F^h. Then each clothing product image in dataset DP is mapped to the codebook in order to obtain its BoVW histogram.

6. EXPERIMENTAL RESULTS

6.1. Implementation

The server stage is implemented in C++ and deployed on a 2.93GHz CPU and 8GB of RAM. A graphical user application is designed for the client side which is implemented in Java and C++ and is deployed for Android smart phones - specifically, we consider the popular Samsung SIII Mini (1GHz dual-core Arm Cortex A9) for demonstration and timing analysis. For demonstration, we design features such as photo querying, viewing top search results, product information (by linking to the retailer's website), and displaying similar products from nearby retailers on a map (refer to Figure 2 for screenshots). Also, products are set to arbitrary locations, whereas for evaluation, we set all products to one retail location so that the more important visual relevance is evaluated.

Several clothing datasets exist but none of them are suitable to evaluate our clothing retrieval task. Datasets mentioned in the current literature either do not solely contain frontal poses [8], or do not feature a large range of clothing and people, or do not feature adults [2], or are low resolution [5] or private. We collect two datasets: a query dataset (DQ) and a product dataset (DP). We primarily consider woman's clothing since it generally exhibits a greater range of colours, textures and shapes than men's and can also be more complex for retrieval than men's due to clothing occlusions by long hair. Dataset DQ consists of a subset of 1000 images from the Fashionista dataset [8] featuring frontal poses suitable for our Viola-Jones face detector.



(a) (b) (c)

Fig. 2: Application: (a) home, (b) search, (c) product map

This dataset contains real world images from a fashion based social network (`chictopia.com`) and is perhaps one of the most challenging for clothing segmentation and retrieval. Since we are concerned with clothing product search, we consider real-world e-commerce images from `esprit.co.uk` for Dataset DP. For this dataset, we collected 1500 images of models in frontal poses wearing woman's tops along with their associated product URLs (so visual retrieval results can link to further details).

6.2. Computational Time

Our system takes on average approximately 6.7 seconds for client processing. Although, we do not fully investigate transmission timing, our system can achieve a total response time of 9 seconds to retrieve results from the server across a 3G data network with excellent smart phone reception. Table 1 lists the computational times of the various stages of the system performed on the client and server. For reliability, the average timings consider a random sample of 10 images with each image in the sample being processed 10 times. These results show that the clothing segmentation is our biggest bottleneck. Our approach is slower than the real time work of [5], however their approach is for a different application, is not implemented in a mobile framework and their dataset is captured on a white background. Our approach is much faster than the work by [8] which works offline on our parent dataset (Fashionista), requiring $2 - 3$GB of memory.

6.3. Accuracy

We select a random sample of 30 images from Dataset A with variation in skin colour and manually segment ground truth to quantitatively analyse our clothing segmentation. Accuracy is reported using the best F-score criterion: $F = 2RP/(P + R)$, where P and R are the precision and recall of pixels in the cloth segment relative to our manually segmented ground

Table 1: Computational Time

Client	Time (ms)
Person Detection	138
Clothing Segmentation	6040
Feature Extraction	411
Feature Quantization	35
Server	Time (ms)
Search and re-ranking	19

truth. We achieve an average F-score over this random sample of 0.857. Since the F-score reaches its best value at 1 and worst at 0, our approach shows reasonable accuracy. Also, this is favourable considering the baseline (GrabCut only) results in an F-score of 0.740 and with the skin elimination routine of Chai rather than our own, 0.808 is achieved. Additionally, by visual inspection of Figure 3, we can see that our approach can segment clothing of persons in various difficult uncontrolled scenes.

Our retrieval results are reported qualitatively in Figure 3. We can see that when the clothing is segmented accurately, the system appears promising with relevant clothing results of a similar colour and shape retrieved. The clothing segmentation stage is important since if it is inaccurate, errors are propagated forward to rest of the system. Segmentation inaccuracies appear to generally be caused by inherent issues such as when scenes contain a garment that is a very similar colour to the background or skin, or there is poor illumination present, or excessive long hair covering the clothing. However, when clothing segmentation fails and our algorithm decides to instead use the unsegmented image to establish features, such as in Figure 3q, we see that the results can still be reasonably relevant although may not be the most accurate.

7. CONCLUSIONS

In this paper, we present a novel mobile client-server framework for automatic visual clothes searching. Our system employs a Bag of Words (BoW) model and proposes an extension of GrabCut for clothing segmentation and a colour descriptor optimized for clothing. We demonstrate a novel application of combining a photo captured on a smart phone (or from social networking) with GPS data to locate clothing of interest at nearby retailers. For future work, we aim to perform a more comprehensive evaluation and integrate more features to train clothing classifiers and re-rank dominant colour results by predicted clothing labels.

8. REFERENCES

[1] eMarketer, "Apparel Drives US Retail Ecommerce Sales Growth," http://www.emarketer.com/newsroom/index.php/apparel-drives-retail-ecommerce-sales-growth, 2012.

[2] A. C. Gallagher and T. Chen, "Clothing cosegmentation for recognizing people," in *CVPR 2008*. IEEE, 2008, pp. 1–8.

[3] B. Hasan and D. Hogg, "Segmentation using Deformable Spatial Priors with Application to Clothing," in *BMVC*, 2010, pp. 1–11.

[4] N. Wang and H. Ai, "Who Blocks Who: Simultaneous Clothing Segmentation for Grouping Images," in *ICCV*, Nov. 2011.

[5] M. Yang and K. Yu, "Real-time clothing recognition in surveillance videos," in *IEEE ICIP*, 2011, pp. 2937–2940.

[6] X. Wang and T. Zhang, "Clothes search in consumer photos via color matching and attribute learning," in *MM*. 2011, pp. 1353–1356, ACM.

[7] X. Chao, M. J. Huiskes, T. Gritti, and C. Ciuhu, "A framework for robust feature selection for real-time fashion style recommendation," in *Workshop on IMCE*. ACM, 2009.

[8] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *CVPR*. IEEE, 2012.

[9] H. Chen, A. Gallagher, and B. Girod, "Describing Clothing by Semantic Attributes," in *ECCV*. 2012, Springer.

[10] J. Sivic, C. L. Zitnick, and R. Szeliski, "Finding people in repeated shots of the same scene," in *BMVC*, 2006, vol. 3, pp. 909–918.

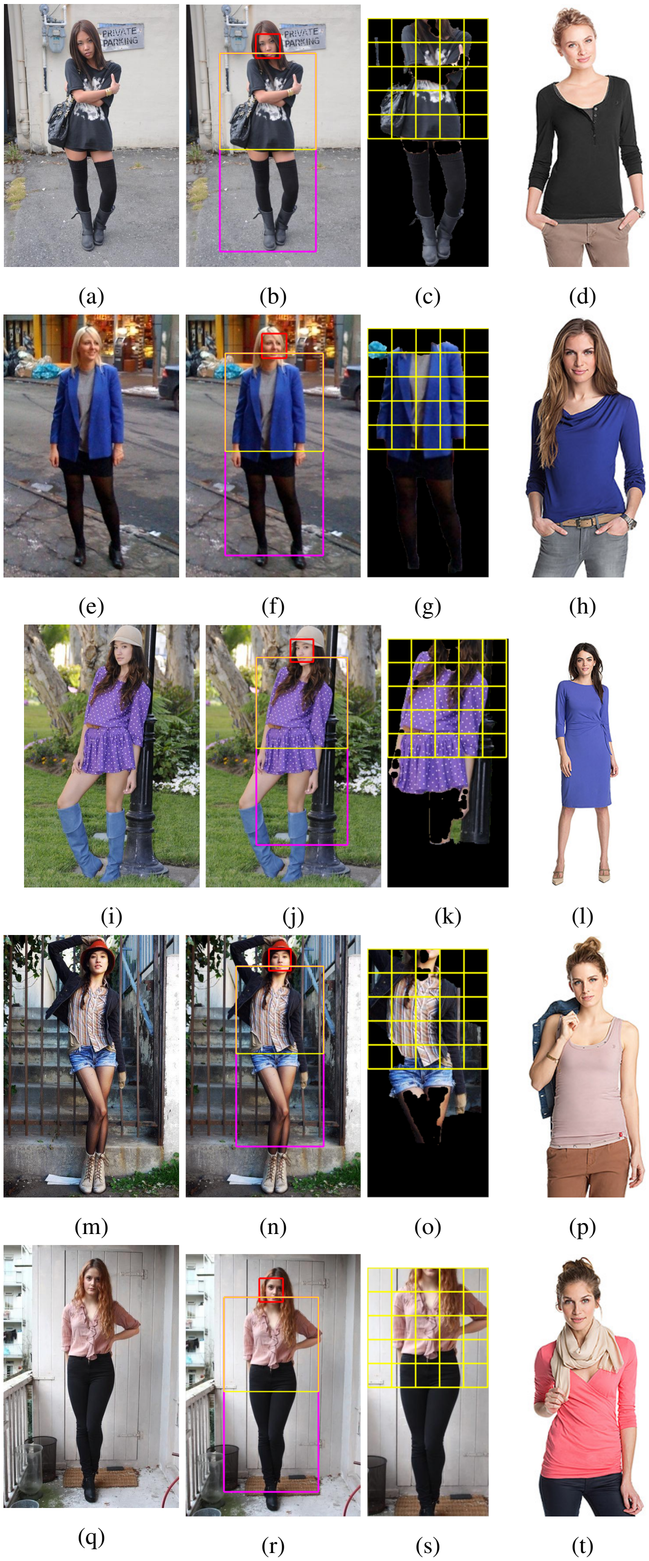
[11] M. W. Lee and I. Cohen, "A model-based approach for estimating human 3D poses in static images," *IEEE TPAMI*, pp. 905–916, 2006.

[12] G. A. Cushen and M. S. Nixon, "Real-Time Semantic Clothing Segmentation," in *ISVC*. 2012, pp. 272–281, Springer.

[13] R. Carsten, K. Vladimir, and B. Andrew, "GrabCut: interactive foreground extraction using iterated graph cuts," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 309–314, Aug. 2004.

[14] D. Chai and K. N. Ngan, "Face segmentation using skin-color map in videophone applications," *CSVT, IEEE Trans on*, vol. 9, no. 4, pp. 551–564, 1999.

[15] T. Sikora, "The MPEG-7 visual standard for content description-an overview," *CSVT, IEEE Trans on*, vol. 11, no. 6, pp. 696–702, 2001.



(a) (b) (c) (d)

(e) (f) (g) (h)

(i) (j) (k) (l)

(m) (n) (o) (p)

(q) (r) (s) (t)

Fig. 3: Qualitative results. Columns depict: (1) query image (I_q), (2) ROI$_p$ (magenta box) and ROI$_u$ (yellow box), (3) segmented clothing (I_c) overlaid with feature extraction grid, and (4) top retrieval candidate $(\hat{j})$.