HUMAN POSTURE RECOGNITION WITH CONVEX PROGRAMMING

Hao Jiang, Ze-Nian Li and Mark S. Drew

School of Computing Science, Simon Fraser University
Burnaby BC, Canada, V5A 1S6
{*hjiangb, li, mark*}*@cs.sfu.ca*

ABSTRACT

We present a novel human posture recognition method using convex programming based matching schemes. Instead of trying to segment the object from the background, we develop a novel multi-stage linear programming scheme to locate the target by searching for the best matching region based on an automatically acquired graph template. The linear programming based visual matching scheme generates relatively dense matching patterns and thus presents a key for robust object matching and human posture recognition. By matching distance transformations of edge maps, the proposed scheme is able to match figures with large appearance changes. We further present object recognition methods based on the similarity of the exemplar with the matching target. The proposed scheme can also be used for recognizing multiple targets in an image. Experiments show promising results for recognizing human postures in cluttered environments.

1. INTRODUCTION

Human posture and activity recognition has attracted a lot of interest in recent years because of its potential important applications in surveillance, human-computer interaction and computer animation. Posture recognition is also one of the most challenging problems in computer vision, because of articulated motion of human bodies and large appearance varieties of clothing. Many traditional human posture recognition systems are based on still cameras and background subtraction; the silhouettes of characters are then used in activity recognition. The difficulty with this scheme is that background subtraction is not robust and not always available, and the method cannot distinguish postures when body parts are covered by silhouettes. One method to solve the problem is by extracting range data for the character in the scene using multiple cameras [1]. But the approach is more expensive to deploy than monocular systems. In this paper, we focus on the problem where only a single camera is available. For monocular systems, matching based methods are promising for solving the problem: we match the features on a template to the object in a target image instead of trying to segment the object from the background. Matching based methods have gained great success in still object recognition [2]. Unfortunately, the distinguished shift invariant features [2] cannot be applied to human posture recognition because they are usually sparse if the object has little texture. Various matching schemes for posture recognition have been presented. A body-part based matching model [3] is presented for human posture recognition. As an extension, an SVM part matching method [4] is further presented. Mori [5] presents a segmentation based approach for part based human posture recognition. The body-part model is flexible, but body parts are usually difficult to match because of their large appearance variation. Another method is to match the target as a whole, e.g. the Chamfer matching based method [6]. One shortcoming of this approach is that it usually needs a lot more templates than part based schemes. Shape matching methods have also been applied for recognition of human actions [7][8]. Shape matching based methods usually need many fewer templates than the Chamfer matching scheme because the template deforms. These schemes work best in relatively clean background settings.

In this paper, we present a scheme to recognize human posture based on matching relatively dense local features. The proposed scheme has the following properties: (1) The method works for cases when reliable background subtraction is unavailable, e.g., for still images; (2) It is quite insensitive to the clothing of the figures in the image; (3) It is able to detect multiple objects in cluttered environments. In this paper, local features are used because they have less variation than the human parts and are therefore more reliable in matching. Unlike global shape features such as shape context [8], local features also enable the proposed scheme to be applicable to matching problems in cluttered environments. In this paper, we propose a novel multi-stage convex programming visual matching scheme. We have proposed a linear programming (LP) formulation [9] and successfully applied the method in motion estimation problems. We have further extended this scheme into a multiple-stage LP relaxation method. The proposed multi-stage relaxation method is found to be more efficient than schemes such as the graph-cut or belief propagation methods for the object matching problem where a large searching range is involved. It can also solve problems for which traditional schemes fail. To suppress the influence of appearance changes for humans, we propose to match the distance transformations of the edge maps of the template and target images. This representation makes matching figures in different clothing possible. We further present a method to quantify the similarity of the template and the target object and form a posture recognition system. The proposed scheme can also be applied for multiple target recognition problems by using a sweeping window approach. Experiments with the proposed scheme show promising results.

2. HUMAN POSTURE ESTIMATION WITH MATCHING

In this section, we present a scheme for estimating the human posture based on convex programming visual matching. First, we present our convex programming matching method, which forms the key component for posture recognition. Then, we study posture recognition based on several similarity measures. Lastly, we extend the proposed method into a multiple-human posture recognition scheme.

2.1. Convex Programming Matching

Mathematically, matching a template to a target image can be formulated as the following energy minimization problem,

$$\min_{u,v} \sum_{(x,y)\in S} C_{x,y,u_{x,y},v_{x,y}} + \sum_{\{(x_1,y_1),(x_2,y_2)\}\in\mathcal{N}} \lambda_{x_1,y_1,x_2,y_2}(|u_{x_1,y_1} - u_{x_2,y_2} - x_1 + x_2| + |v_{x_1,y_1} - v_{x_2,y_2} - y_1 + y_2|)$$

We need to match each point $(x, y) \in S$ in a template to a point $(u_{x,y}, v_{x,y})$ in the target image such that the energy is minimized. $C_{x,y,u,v}$ is the matching cost; $\mathcal{N}$ is the set of neighboring sites; the second term in the energy function is a penalty term which smooths the mapping of neighbor sites in S; $\lambda_{x_1,y_1,x_2,y_2}$ are smoothing factors. Here we assume that S is a finite set.

We propose a multiple-step convex programming method to solve the non-linear optimization problem, in which we relax the non-linear problem into a sequence of linear programming problems based on the previous LP relaxation solution and gradually shrink the searching region. By solving a sequence of LP approximations of the non-linear programming problem, the result of optimization is greatly improved compared to a single-step linear programming method. At stage n, $n = 0..N$, the linear programming relaxation is

$$LP_n : \min \sum_{(i,j)\in S,(p,q)\in\mathcal{F}^n_{i,j}} C_{i,j,p,q}\xi_{i,j,p,q} + \sum_{\{(i,j),(k,l)\}\in\mathcal{N}} \lambda_{i,j,k,l}(u^+_{i,j,k,l} + u^-_{i,j,k,l} + v^+_{i,j,k,l} + v^-_{i,j,k,l})$$

subject to:

$$\sum_{(p,q)\in\mathcal{F}^n_{i,j}} \xi_{i,j,p,q} = 1, \forall (i,j) \in S$$

$$\sum_{(p,q)\in\mathcal{F}^n_{i,j}} p\xi_{i,j,p,q} = u^n_{i,j}, \forall (i,j) \in S$$

$$\sum_{(p,q)\in\mathcal{F}^n_{i,j}} q\xi_{i,j,p,q} = v^n_{i,j}, \forall (i,j) \in S$$

$$u^n_{i,j} - u^n_{k,l} - i + k = u^+_{i,j,k,l} - u^-_{i,j,k,l}, \forall \{(i,j),(k,l)\} \in \mathcal{N}$$

$$v^n_{i,j} - v^n_{k,l} - j + l = v^+_{i,j,k,l} - v^-_{i,j,k,l}, \forall \{(i,j),(k,l)\} \in \mathcal{N}$$

with bounds:

$$\xi_{i,j,p,q} \geq 0, u^+_{i,j,k,l}, u^-_{i,j,k,l}, v^+_{i,j,k,l}, v^-_{i,j,k,l} \geq 0$$

where $\mathcal{F}^n_{i,j}$ is the target set for source point (i, j), in the matching region $\mathcal{R}^n_{i,j}$. In fact, we do not have to include the whole set of candidate matching costs in the LP relaxation [9]. It is not difficult to prove that the LP relaxation is equivalent to reformulating the original non-linear problem by approximating $C_{i,j,p,q}$ by its lower convex hull in $\mathcal{R}^n_{i,j}$, for each $(i, j) \in S$. Therefore we can replace $\mathcal{F}^n_{i,j}$ by the set of vertex coordinates of the lower convex hull of $\{C_{i,j,p,q}, \forall (p, q) \in \mathcal{F}^n_{i,j}\}$. The search regions $\mathcal{R}^0_{i,j}$ for LP$_0$ are the whole target image. We update the searching region of stage n, $n \geq 1$, by keeping $(u^{n-1}_{i,j}, v^{n-1}_{i,j})$ inside the search region and moving the four region boundaries inward. If $(u^{n-1}_{i,j}, v^{n-1}_{i,j})$ falls on the region boundary, we move the other boundaries inward. We applied a revised simplex method to solve the LP problem. The simplex method has been found to be very efficient in applications even though the worse case complexity is exponential. If we ignore the complexity of finding the lower convex hull, an estimate of the average complexity of the proposed matching scheme is $O(|S| \cdot |\mathcal{F}|^{1/2} \cdot (\log |\mathcal{F}| + \log |S|))$. Experiments also confirm that the average complexity of the proposed optimization scheme increases more slowly with the searching window size than previous methods such as the graph cut scheme, whose average complexity is linear with respect to $|\mathcal{F}|$.

For posture recognition problems, the features selected for the matching process have to be insensitive to the appearance changes of human objects. The edge map contains all the shape information of an object, and at the same time removes the difference due to color changes. The edge feature has been widely applied in Chamfer matching schemes. One problem of traditional Chamfer matching is that it does not take into consideration the directional information of the edges. And, extracting directional information is usually a difficult problem since the local orientation has a great deal of ambiguity at positions such as corners. To overcome this problem, we propose the use of small blocks, centered on the edge pixels, of the *distance transform* of an image's edge map, as the matching feature. A distance transform converts a binary edge map into its corresponding grayscale representation, where the intensity of a pixel is proportional to its distance to the nearest edge pixel. Denoting the square block of the distance transform of I's edge map centered at the edge pixel (i, j) as $\mathbf{d}_{i,j}(I)$, the cost of matching is defined as

$$C_{i,j,p,q} = \frac{1}{\Delta^2\sqrt{\sigma_s\sigma_t}}||\mathbf{d}_{i,j}(I_s) - \mathbf{d}_{p,q}(I_t)||$$

where I_s and I_t are the template and target images respectively; $||.||$ is the cityblock norm in this paper; σ_s and σ_t are the standard deviations of $\mathbf{d}_{i,j}(I_s)$ and $\mathbf{d}_{p,q}(I_t)$ respectively; Δ is the size of the square block. The orientation information is now integrated in the proposed feature. For instance, there is now a big difference for two features on orthogonal edges. By using a small window that does not cover multiple edges, the proposed feature is also scale invariant. In this paper, the features are randomly selected on the edges of the template. The neighboring relation $\mathcal{N}$ is defined by the edges of the graph generated by Delaunay triangulation of the feature points on the template.

2.2. Similarity Measures

After finding the matches of the feature points in the template with corresponding points in the target image based on the proposed method, we need further to decide how similar these two constellations of matched points are and whether the matching result corresponds to the same event as in the exemplar. We use the following quantities to measure the difference between the template and the matching object. The first measure is D, defined as the average of pairwise length changes from the template to the target. To compensate for the global deformation, a global affine transform $\mathcal{A}$ is first estimated based on the matching and then applied to the template points before calculating D.

$$D = \frac{1}{|\mathcal{N}|} \sum_{\{(i,j),(k,l)\} \in \mathcal{N}} ||\mathcal{A}\circ(i-k, j-l) - (\hat{u}_{i,j} - \hat{u}_{k,l}, \hat{v}_{i,j} - \hat{v}_{k,l})||$$

where $(\hat{u}_{i,j}, \hat{v}_{i,j})$ is the matching point of (i,j), from the LP scheme. The second measure is the average matching cost M.

$$M = \frac{1}{|S|} \sum_{(i,j)\in S} C_{i,j,\hat{u}_{i,j},\hat{v}_{i,j}}$$

Shape context [8] is also found to be useful in classifying objects. To remove the interference of background clutter, we calculate a prediction object mask based on the predefined template mask. The prediction is based on the affine transformation $\mathcal{A}$ estimated from the matching points. Only the edge points inside the object mask and prediction mask are used to calculate the shape context. Similarly, $\mathcal{A}$ is first applied to the template before calculating the shape context. Let $\mathbf{h}_{i,j}(I)$ be the histogram in the polar space centered at (i,j) of image I. The difference of shape context H is defined as

$$H = \frac{\sum_{(i,j)\in S} ||\mathbf{h}_{\mathcal{A}\circ(i,j)}(\mathcal{A}\circ E(I_s)) - \mathbf{h}_{(\hat{u}_{i,j},\hat{v}_{i,j})}(E(I_t))||}{|S||E(I_s)|}$$

where $E(I)$ is the edge point set of image I. These features can be fed into a *linear classifier* to detect whether the object appears in the target window. The linear combination of the three features forms a matching score. We define that an event occurs in the target window if the matching score is lower than a threshold. Experiments show that only about 100 randomly selected feature points are needed in calculating D, M and H.

2.3. Dealing with Multiple Targets

If there is only one human object in the image, we can directly use the above method to locate the object and decide on the corresponding posture by comparing with the exemplars. If there is no knowledge about the number of human objects in the image, we use the scheme of overlapping detection windows. We sweep a window across the image; the window size is equal to or a little larger than the template size. For each window we apply the proposed matching scheme to detect whether there is a specific object in the image. If there is a target in the window, the matching object is kept for further verification. Clearly, if hardware permits, parallel processing can be applied for each detection window.

3. EXPERIMENTAL RESULTS

Fig. 1 shows the advantage of using our deformable matching scheme when we only have one template available. As shown in the example, small appearance changes between the target and the template will result in the failure of the Chamfer matching method. Greedy schemes also meet with great difficulty since there are a lot of ambiguities in matching distance transformation images. The proposed LP based method can solve the problem.

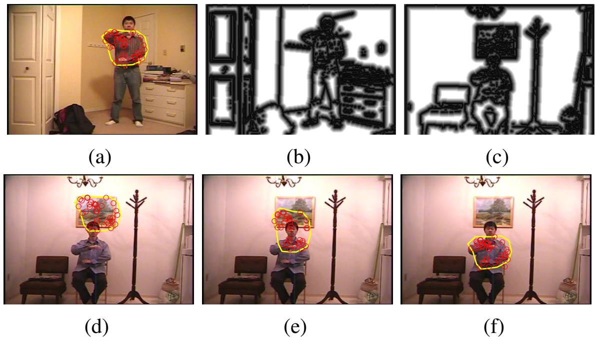


Fig. 1. An example where Chamfer matching fails. (a): Template; (b, c): Source and target distance transforms; (c): Chamfer matching result; (c): Iterative Conditional Modes matching result; (d): Our LP matching result.

In another experiment, we study the following retrieval problem: we use a template image and retrieve the best matches in an image data set. The data set is extracted from a video sequence. Repeated postures are manually removed from the data set. In this experiment, there is only one human object in the image. The search range is the whole target image. The character in the image has very different clothing in the template and target image. The scoring function is a linear combination of the proposed three similarity measures and trained as a linear classifier by another data set. Here a lower score indicates a better match. Fig. 2 shows retrieval results for 4 different postures. The first two best matches are shown. As shown in these experiments, the proposed method still works well in the challenging case when the arms are in front of the torso. Fig. 3 shows an activity retrieval result, in which different characters are involved. The best matches are retrieved from an 800-frame video by searching every other frame. The recall and the precision for the first 100 matches are 83% and 91% respectively.

We also apply the proposed scheme for recognizing a walking person in the image, based on the sweeping window scheme. Some results are shown in Fig. 4. A single template

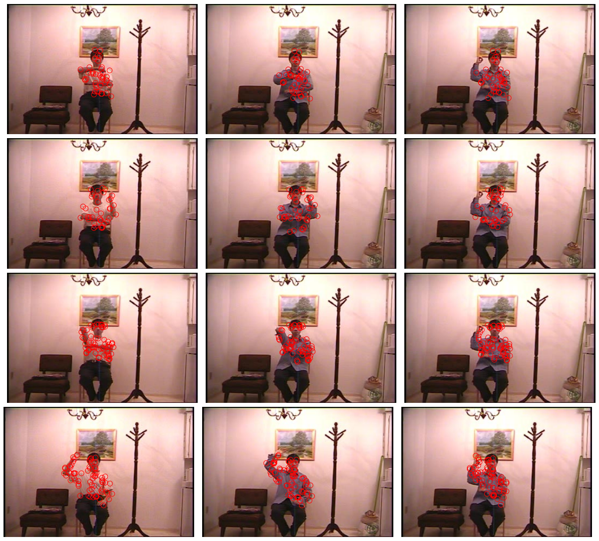


Fig. 2. Left column: Templates; Middle and right column: Best and next-best matches.

is used. The matching score is compared with a fixed threshold to determine whether an object is in the scene. Closely overlapping detection results are automatically removed by keeping only the best-matching one.

For still images, false alarms and mismatching are still problems difficult to avoid. For videos, background subtraction could remove part of interference factors and thus further improves the detection result.

4. CONCLUSION

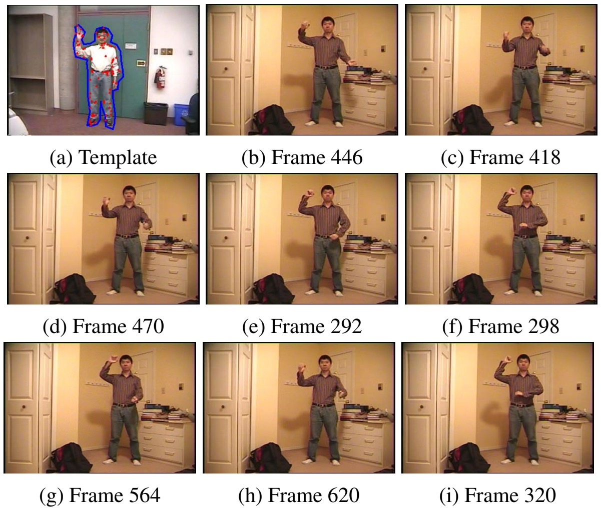
We propose a novel multiple-step relaxation linear programming method which is more efficient than schemes such as the graph-cut or belief propagation methods for the object matching problem where a large searching range is involved. It can also solve problems for which other schemes fail. As well, we propose using the distance transformations of the edge maps to match the template and target images. This representation makes matching of different objects in the same class possible. Experiments show promising results for human activity detection in cluttered environments. In future work, the postures can be linked to some possible activities of the character in the image such as "walking", "running", "dancing" thus forming an activity retrieval system. The proposed scheme can also be directly applied for general object recognition problems.

5. REFERENCES

[1] K.M.G. Cheung, S. Baker, T. Kanade, "Shape-from-silhouette of articulated objects and its use for human body kinematics estimation and motion capture", CVPR, pp. I:77-84 vol.1, 2003.

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints", International Journal of Computer Vision, 60, pp. 91-110, 2004.

[3] P.F. Felzenszwalb, D.P. Huttenlocher, Efficient matching of pictorial structures, CVPR, pp. II:66-73 vol.2, 2000.

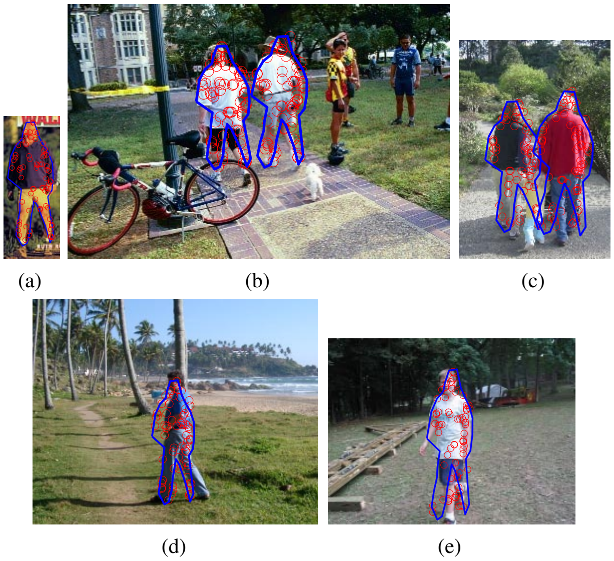


(a) Template (b) Frame 446 (c) Frame 418

(d) Frame 470 (e) Frame 292 (f) Frame 298

(g) Frame 564 (h) Frame 620 (i) Frame 320

Fig. 3. (a): Template; (b, c, d, e, f, g, h, i): Eight best match frames.



(a) (b) (c)

(d) (e)

Fig. 4. (a): Template; (b, c, d, e): Detection results.

[4] R. Ronfard, C. Schmid, and B. Triggs, "Learning to Parse Pictures of People", ECCV, LNCS 2353, pp. 700–714, 2002.

[5] G. Mori, X. Ren, A. Efros, and J. Malik, "Recovering human body configurations: combining segmentation and recognition", CVPR, pp.II:326-333, 2004

[6] D. M. Gavrila and V. Philomin, "Real-time object detection for smart vehicles", ICCV, pp. 87-93, Kerkyra, Greece, 1999.

[7] S. Carlsson and J. Sullivan, "Action recognition by shape matching to key frames", IEEE Computer Society Workshop on Models versus Exemplars in Computer Vision, 2001.

[8] G. Mori and J. Malik, "Estimating human body configurations using shape context matching", ECCV, LNCS 2352, pp. 666–680, 2002.

[9] H. Jiang, Z.N. Li, and M.S. Drew, "Optimizing motion estimation with linear programming and detail-preserving variational method", CVPR, pp.I:738-745, 2004.