TRACKING POINTS ON DEFORMABLE OBJECTS WITH RANKLETS

F. Smeraldi, A. Del Bue and L. Agapito

Department of Computer Science
Queen Mary University of London
London, E1 4NS, U.K.

ABSTRACT

We present a robust algorithm for point tracking on deformable objects. The key elements are the use of orientation selective rank features (*ranklets*), local filter adaptation and dynamic model update. A multi-scale vector of ranklets is used to encode a neighbourhood of each tracked point. The shape of the filters is optimised for each neighbourhood independently. Substantial appearance variations are catered for by maintaining a stack of models for each tracked point. This enables the system to recalibrate whenever the object reverts to its original appearance.

1. INTRODUCTION

Point tracking is a difficult problem due to the necessity of having a highly discriminative feature representation that can match the location of the tracking targets precisely, while at the same time allowing for sufficient flexibility to compensate for the appearance variation of the tracked objects in time.

Correlation–based algorithms are particularly dependent on brightness constancy [1, 2]; to overcome this limitation, the use of explicit photometric models has been proposed in [3]. Feature–based approaches based on Gabor filters [4, 5] or other orientation–selective features [6, 7] benefit from increased invariance because of the inherent robustness of orientation information. The use of such features for tracking is supported the ubiquitous presence of orientation selectivity in biological vision systems [8]. However, the highest degree of invariance to illumination changes is provided by rank features, that have been widely applied in the related problem of stereo correspondence (see for instance [9]). For these reasons, orientation selective rank features would appear to be well suited for point tracking.

In this paper we present a tracking algorithm based on *ranklets*, a recently developed family of multi-scale rank features that unite an orientation selectivity pattern similar to Haar wavelets with the wide degree of invariance afforded by rank features [10].

Similarly to classic multi-scale algorithms [4, 7], our approach uses a vector of ranklets to encode an appearance based description of the neighbourhood of each of a sparse set of tracked points. The aspect ratio of the filters is adapted to each local neighbourhood to maximise filter response; this is expected to make the representation locally more discriminative, thus minimising drift.

Some model update mechanism is generally required to compensate for appearance variations of the tracked targets in time [11, 12]. In our case, the rank based representation affords a degree of robustness to small variations that alleviates this problem. Larger variations arising from rotations or deformations of the object are handled in a straightforward but flexible way by maintaining a dynamic stack of models for each tracked point (an approach related to the ideas in [13]). This allows the tracker to follow the points through a wide range of deformations with limited drift, and eventually recalibrate whenever the object reverts to its original appearance.

The use of orientation selective rank features, filter adaptation and dynamic model updating makes our algorithm particularly suitable for tracking points on deformable structures. We present experimental results over a sequence of a face performing substantial three-dimensional rotations and extreme expression changes.

2. ORIENTATION SELECTIVE RANK FEATURES

Before introducing the details of our algorithm, we briefly summarise the definition of ranklets (we refer the reader to [10] for a more complete treatment).

Ranklets are a family of orientation selective rank features designed in close analogy with Haar wavelets. However, whereas Haar wavelets are a set of filters that act linearly on the intensity values of the image, ranklets are defined in terms of the relative order of pixel intensities.

Consider the three Haar wavelets $h_i(\mathbf{x}), i = 1,2,3$ shown in Figure 1, supported on a local window S. The aim with ranklets is to perform a nonparametric comparison of the relative brightness of the pairs of pixel sets $\mathsf{T}_i = h_i^{-1}(\{+1\})$ and $\mathsf{C}_i = h_i^{-1}(\{-1\})$.

A straightforward nonparametric measure of the intensity of the pixels in T_i compared to those in C_i can be ob-

0-7803-9134-9/05/$20.00 ©2005 IEEE

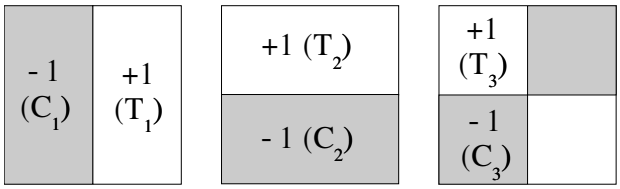

Fig. 1. The three two-dimensional Haar wavelets $h_1(\mathbf{x})$, $h_2(\mathbf{x})$ and $h_3(\mathbf{x})$ (from left to right). Letters in parentheses refer to the T and C pixel sets defined in the text.

tained as follows. Consider the set $\mathsf{T}_i \times \mathsf{C}_i$ of all pixel pairs $(\mathbf{x},\mathbf{y})$ with $\mathbf{x} \in \mathsf{T}_i$ and $\mathbf{y} \in \mathsf{C}_i$. Let $\mathcal{W}_{YX}^i$ be the number of such pairs in which the pixel from the set T_i is brighter than the one from C_i, that is

$$\mathcal{W}_{YX}^i = \#\{(\mathbf{x},\mathbf{y}) \in \mathsf{T}_i \times \mathsf{C}_i | I(\mathbf{x}) > I(\mathbf{y})\} \qquad (1)$$

(the reason behind the fancy notation will become clear later on). Essentially, $\mathcal{W}_{YX}^i$ will be close to its maximum value, i.e. the number of pairs in $\mathsf{T}_i \times \mathsf{C}_i$, if the pixels in the T_i region are brighter than those in the C_i region; conversely, it will be close to its minimum value (i.e. 0) if the opposite is true. Remembering that the T_i and C_i sets coincide by definition with the "+1" and "-1" regions of the wavelets in Figure 1, we see that each $\mathcal{W}_{YX}^i$ displays the same orientation selective response pattern as the corresponding Haar wavelet h_i.

The procedure outlined above can be carried out with complexity of at most $O(N \log N)$, where N is the number of pixels in the support window S. For this, it will suffice to sort the pixels $\mathbf{x} \in \mathsf{S}$ according to their intensity $I(\mathbf{x})$. Indicate the rank of pixel $\mathbf{x}$ with $\pi(\mathbf{x})$; we then have

$$\mathcal{W}_{YX}^i = \sum_{\mathbf{x} \in \mathsf{T}_i} \pi(\mathbf{x}) - (N/2+1)N/4 \qquad (2)$$

(for a proof, see [14]). The quantity $\mathcal{W}_{YX}^i$ is known as the Mann-Whitney statistics for the observables (the pixels) in T_i and C_i (according to the standard terminology, these would be the "Treatment" and "Control" sets). The Mann-Whitney statistics is equivalent to the Wilcoxon statistics $\mathcal{W}_s$ [14].

For practical reasons, it is convenient to define ranklets as

$$\mathcal{R}^i = 2\frac{\mathcal{W}_{YX}^i}{N^2/4} - 1, \qquad (3)$$

so that their value increases from -1 to $+1$ as the pixels in T_i become brighter than those in C_i.

3. ADAPTIVE APPEARANCE–BASED MODELLING

In analogy with classic approaches involving multi-scale features [4, 7], we choose to encode the local image neighbourhood of each point by means of a vector of ranklets consisting of a total of 9 filters arranged in 3 frequency channels and 3 orientation channels (corresponding, as in the case of Haar wavelets, to horizontal, vertical and diagonal edges).

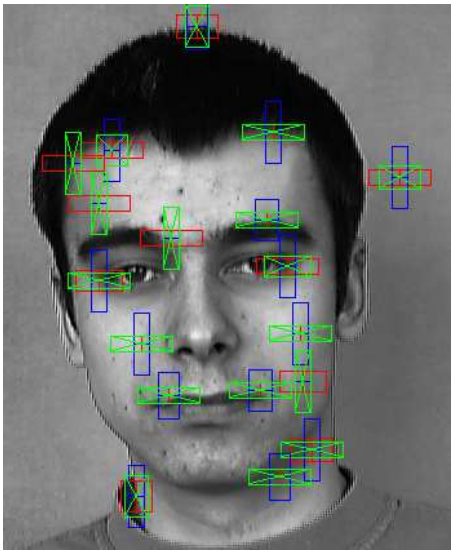

Fig. 2. Results of filter adaptation for a few tracked points (lowest frequency channel). The rectangles represent the support of the filters. The orientation selectivity of each filter is indicated by vertical ($\mathcal{R}^1$), horizontal ($\mathcal{R}^2$) or diagonal ($\mathcal{R}^3$) lines drawn across the corresponding rectangle.

The target points for tracking are selected based on a measure of saliency that is proportional, for each point, to the norm of the corresponding ranklet vector (see the automatically selected points in Figure 2).

For each tracked point, an optimisation step is performed to adapt the shape of the filters to the specific appearance of the neighbourhood of the point. This is done by independently varying the aspect ratio of the support of each ranklet in order to obtain the largest possible response (the area of the support is kept constant in the process). The purpose of adaptation step is maximising the saliency of the tracked location across the local neighbourhood, thus facilitating tracking. The support of a few adapted filters is shown in Figure 2.

Tracking is performed by using, for each tracked point, the adapted ranklet vector as a model. In each subsequent frame, a gradient descent algorithm is employed in a neighbourhood of the previous position of the point in order to find the location that gives the best match.

4. UPDATING THE MODELS

Due to the deformations and pose changes of the tracked object, the quality of the match between the features extracted at the location of each point and the corresponding model generally deteriorates with time. This eventually results in failure to track the points.

To prevent this, we maintain a model stack for each tracked point. A new model is acquired from the current best estimate of the position of a point whenever the residual distance from the original model (after matching) exceeds a threshold τ. The filter adaptation step is repeated and the new model is stored on the stack above the previous one. This procedure is repeated when necessary up to a given maximum number of models, after which the particular point is considered lost.

While tracking each point the most recently acquired model, which is on top of the stack, is used first. A further gradient descent is then initiated in an attempt to match the previous model on the stack. If the resulting discrepancy is now below τ the last model is discarded by popping the stack, and the point is assumed to have recovered the appearance it had at an earlier time. The algorithm then attempts to work its way further down the stack by iterating the procedure. In this way, the active model is always the oldest acquired model for which the matching distance does not exceed τ.

The model stack provides a mechanism for tracking a point across a range of deformations and pose variations, during which the point may occasionally revert to its original appearance. Upon creation of a new model, an added check is performed to allow "grafting" it next to the most similar model already present in the stack (if this is different from the active model). The contents of the stack above the grafting position are then discarded. Thus a point is not required to return to its original appearance by going through the same series of deformations in reverse order (although this will often be the case, for instance, for the points of a face that is rotating left to right and then right to left). Points are discarded when thresholds for the maximum number of models or the maximum frame-to-frame drift are exceeded.

5. EXPERIMENTAL RESULTS

We have tested our tracking algorithm on a 1075 frame sequence showing the face of a subject performing substantial 3D rotations and extreme changes of expression.

A total of 91 points were initialised automatically according to the saliency criterion described in Section 3. The result of the filter adaptation step on the first frame is shown in Figure 2. As can be seen, the filters deform in such a way as to maximise the contrast between their T and C areas. In the case of the diagonal filters ($\mathcal{R}^3$, see Section 2), this has the effect of tuning the filter to the locally predominant orientation. As Figure 3 shows, the tracker was able to follow a good number of points reliably throughout the sequence, even in relatively poorly textured areas such as the cheekbones. Throughout the 1075 frame sequence only 5 points out of the initial 91 were lost, showing that the tracker can cope with significant deformations and pose changes. Two of these points were unstable and were lost at an early stage (before frame 22), while the rest were eliminated by the model stack overflow condition. In our experiments we allowed a maximum of 20 models for each point.

Figure 3 also describes the operation of the model update mechanism. For each frame, a histogram of the depth of the model stack across all tracked points is presented. As can be seen, more models are acquired for a large number of points in the presence of large deformations. The points then revert to the original models as the subject recovers its original appearance. The combined effect of the model updating technique and backtracking phases is to allow the tracker to follow the points through a wide range of deformations, while at the same time limiting drift. However a certain number of unreliable points remains, in correspondence of poorly textured areas.

A direct comparison between existing algorithms and our own is inherently difficult since the specific saliency measure used by each tracker would typically result on each one being initialised over a different set of points. A qualitative analysis shows that the classical KLT (Kanade-Lucas-Tomasi) tracker [1] looses track of a high percentage of the points after a few frames on the same sequence as above.

6. CONCLUSIONS

We have presented a feature–based, adaptive point tracking algorithm based on orientation selective rank features. A vector of ranklets is used to model a neighbourhood of each tracked point. As with all other rank features, the representation has a high degree of invariance to the illumination and appearance changes resulting from 3D rotations and deformations of the object. In addition, the orientation selectivity of ranklets adds to the discriminative power of the representation, thus minimising point drift. This effect is enhanced by locally adapting the aspect ratio of the filters to the appearance of the tracked points, thus making the feature extraction process point–specific.

Large variations in appearance are handled by a model stack mechanism, which effectively keeps track of the feature space trajectory of the single points. By adaptively storing new models on the stack when deformations make it necessary and reverting to older models when the quality of the match allows it, the algorithm is able to cope with significant variations without loosing track. Precision is gradually recovered as the points revert to their original appearance.

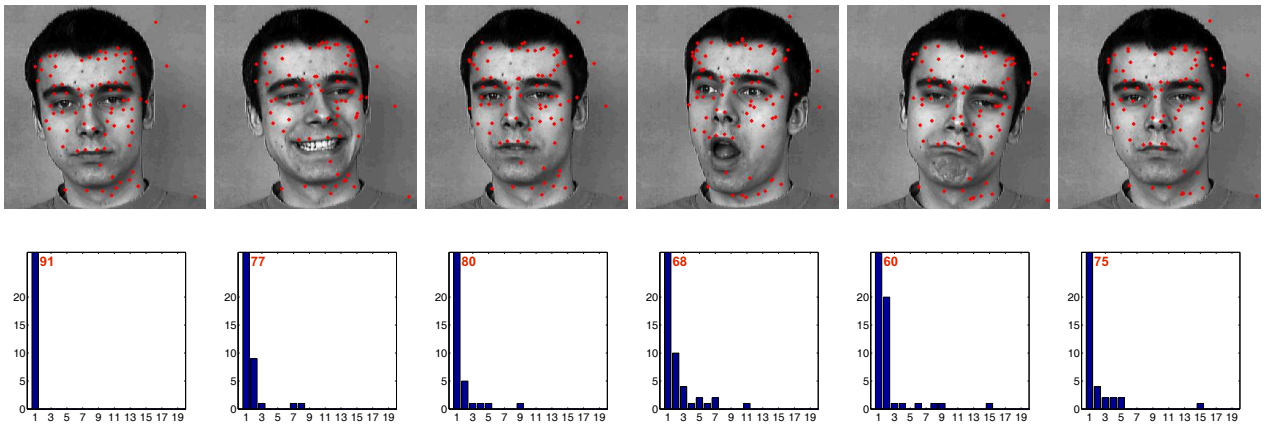

Fig. 3. *Results for the new tracker on a sequence where the subject performed rigid and non-rigid motion. The histograms show how many points (y axis) have how many models (x axis) on their model stack. Notice how new models are added to accommodate large deformations (Frame 506 and 925); by Frame 1075 most points have reverted to their original model*

In our experiments, the tracking algorithm has showed its ability to handle the deformations and pose changes of a non-rigid object, such as a face, effectively. Our current work concentrates on the estimation of 3D shape and modes of deformation of the tracked object [15], which will in turn impose constraints on the image motion of the points.

Acknowledgements

The authors would like to thank L. Zalewski who provided the sequences. ADB holds a Queen Mary Studentship. Part of this work was supported by EPSRC Grant GR/S61539/01.

7. REFERENCES

[1] Jianbo Shi and Carlo Tomasi, "Good features to track," in *IEEE Conference on Computer Vision and Pattern Recognition, Seattle (USA)*, June 1994, pp. 593–600.

[2] A. Fusiello, E. Trucco, T. Tommasini, and V. Roberto, "Improving feature tracking with robust statistics," *PAA*, vol. 2, no. 4, pp. 312–320, 1999.

[3] H. Jin, P. Favaro, and S. Soatto, "Real–time feature tracking and outlier rejection with changes in illumination," in *Proc. of ICCV*, July 2001, pp. 684–689.

[4] B. S. Manjunath, C. Shekhar, and R. Chellappa, "A new approach to image feature detection with applications," *Pattern Recognition*, vol. 31, pp. 627–640, 1996.

[5] J. Wieghardt, R. P. Würtz, and C. v.d. Malsburg, "Gabor–based feature point tracking with automatically learned constraints," in *Dynamic Perception*, R. P. Würtz and M. Lappe, Eds. 2002, pp. 121–126, Infix Verlag.

[6] J. Bigun, T. Bigun, and K. Nilsson, "Recognition by symmetry derivatives and the generalized structure tensor," *IEEE–PAMI*, vol. 26, pp. 1590–1605, 2004.

[7] R. P. N. Rao and D. H. Ballard, "An active vision architeture based on iconic representations," *Artificial Intelligence Journal*, vol. 78, pp. 461–505, 1995.

[8] J. G. Daugman, "Two-dimensional spectral analysis of cortical receptive field profiles," *Vision Research*, vol. 20, pp. 847–856, 1980.

[9] R. Zabih and J. Woodfill, "Non-parametric local transforms for computing visual correspondence," in *Proceedings of the 3rd ECCV*, 1994, pp. 151–158.

[10] F. Smeraldi, "Ranklets: orientation selective non-parametric features applied to face detection," in *Proc. of ICPR*, 2002, vol. 3, pp. 379–382.

[11] M. Black and A. Jepson, "Eigen–tracking: Robust matching and tracking of articulated objects using a view–based representation," *IJCV*, vol. 36, no. 2, pp. 63–84, 1998.

[12] G. D. Hager and P. N. Belhumeur, "Efficient region tracking with parametric models of geometry and illumination," *IEEE Transactions on PAMI*, vol. 20, no. 10, pp. 1025–1039, 1998.

[13] I. Matthews, T. Ishikawa, and S. Baker, "The template update problem," in *Proceedings of the British Machine Vision Conference, Norwich, UK*, R. Harvey and J. A. Bangham, Eds., 2003, vol. II, pp. 649–658.

[14] E. L. Lehmann, *Nonparametrics: Statistical methods based on ranks*, Holden-Day, 1975.

[15] A. Del Bue, F. Smeraldi, and L. Agapito, "Non-rigid structure from motion using ranklet-based tracking and non-linear optimization," *Image and Vision Computing*, 2005, Accepted for publication.