ELSEVIER

Image and Vision Computing 20 (2002) 249–256

image AND vision COMPUTING

www.elsevier.com/locate/imavis

How to measure the pose robustness of object views

Gabriele Peters[a,*], Barbara Zitova[b], Christoph von der Malsburg[a,1]

[a]*Institut für Neuroinformatik, Systembiophysik, Ruhr-Universität Bochum, Universitätsstr. 150, D-44780 Bochum, Germany*
[b]*Department of Image Processing, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Pod vodárenskou věží 4, 182 08 Praha 8, Czech Republic*

Received 16 October 2000; accepted 18 December 2001

Abstract

The viewing hemisphere of a three-dimensional object can be partitioned into areas of similar views, which provide pose robustness. We compare two procedures for measuring the robustness of views to pose variation: *tracking* of object features, i.e. Gabor wavelet responses, by utilizing the continuity of successive views and *matching* of features in different views, which are assumed to be independent. Both procedures proved to be appropriate to detect canonical views. We found no difference concerning the size of the view bubbles, but tracking provides more precise correspondences than matching. Tracking is more appropriate for recognizing *changes* of features, whereas matching is more suitable if features of the *same* appearance are to be found.

Keywords: Three-dimensional object perception; Pose robustness; Matching/tracking object features; Canonical views

1. Subject of investigation

Many models have been proposed for three-dimensional object perception. Besides volume-based object representations, which seem to be very economical but often require the interaction from a user to acquire them, as for example, described in Ref. [1], many computational models combine two-dimensional views into the equivalent of a three-dimensional object representation. Examples are the manifold approach applied [2–4] and the recognition of three-dimensional objects utilizing support vector machines [5,6].

Among the different models for three-dimensional object perception, the notion of a *canonical view* is a prominent topic. It can be regarded as a view which is easier to recognize than other views of the same object. A hard definition does not exist, even its properties are controversial. Palmer et al. [7] describe canonical views as the ones that "humans find easiest to recognize and regard as most typical". Open questions concerning canonical views are the number of views necessary for different visual tasks and their statistical distribution on the viewing sphere. Malik and Whangbo [8], for instance, have demonstrated that a uniform distribution is inappropriate. Weinshall and Werman [9] have shown that the likelihood to observe a certain view of an object correlates with the view's robustness against pose variation, i.e. how little the image changes when the viewpoint is slightly changed. The most likely views are often the 'flattest' views of an object.

For pose-invariant object recognition and pose estimation of objects, it is necessary to utilize an appropriate object representation. An obvious but naive representation might consist of densely spaced views of an object's viewing sphere. Our aim is to reduce such a 'full' representation to only some representative views and the relations between them. Such a sparse representation belongs to the *aspect graph* approaches proposed in Refs. [10,11]. To choose representative views (*aspects*) for a final representation, our plan is first to determine for each view of a full representation a surrounding area of similar views, termed *view bubble*, which provides robustness to pose variation (see Fig. 1). Later, the aspects for the final representation of the object can be derived from the overlaps of the view bubbles.

In this paper we describe the generation of the view bubbles. We restrict our investigation to the upper hemisphere of an object's viewing sphere. We compare two procedures of determining the similarity of two views: *matching* a representing graph of one view to another view, and *tracking* object features, i.e. Gabor wavelet responses, from one view to another. During the matching procedure, each view is treated independently, whereas the tracking procedure utilizes the continuity of neighboring views. Our investigations were guided by the question which procedure—matching or tracking—is more appropriate to find for each view of

* Corresponding author. Tel.: +49-234-32-27971; fax: +49-234-32-14210.
E-mail address: peters@1s7.cs.uni-dortmund.de (G. Peters).
[1] Tel.: +49-234-32-27997; fax: +49-234-32-14210.

0262-8856/02/$ - see front matter
PII: S0262-8856(02)00006-9

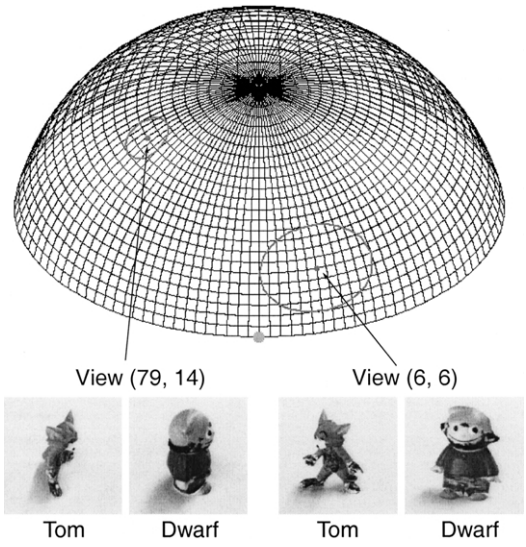

Fig. 1. Viewing hemisphere with examples of view bubbles. The representation of a viewing hemisphere consists of 100×25 views. Each crossing of the grid stands for one view. The angle between two neighboring views is 3.6° in either direction. The dot in front marks view (0,0). The depicted view bubbles have been determined with the tracking procedure for object Tom. View (79,14) provides a small view bubble. It includes views which cover a range of 21.6° in *x*-direction and 14.4° in *y*-direction. View (6,6) provides a larger view bubble, which covers a range of 43.2 and 28.8°, respectively.

the hemisphere view bubbles of *maximal size* containing views of *maximal similarity*?

2. Description of the system

2.1. Preprocessing

For each recorded view of an object, we first perform a segmentation based on gray levels which separates the object from the background. Then we put a grid graph onto the segment of the image which has been assigned to the object. At each vertex of the graph, we extract features which describe the surroundings of the vertex, i.e. local features of the special view of the object. Thus, we derive a representation for each view in the form of a *model graph* which provides the basis of both, the matching and the tracking procedure (see Fig. 2).

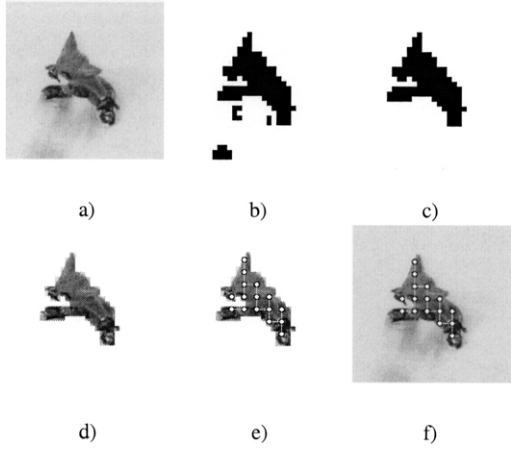

Fig. 2. Preprocessing: (a) original image, (b) gray level segmentation, (c) centered segmentation after eliminating wrong segments, (d) original image masked with the result of centered segmentation, (e) grid graph on object segment, and (f) grid graph on original image.

Segmentation: The segmentation method is based on the system in Ref. [12]. An image is divided into small patches. Each patch receives a label that encodes its membership of one of the several segments. We need two segments because we want to separate one object from the background. The aim is to find the label configuration which encodes the 'correct' segmentation of the given scene. Each label interacts with neighboring labels via an interaction matrix. The interaction between two labels is computed from the difference in mean intensity at the corresponding image regions. The desired segmentation results from coarse-to-fine dynamics which relax to a local energy minimum. The segmentation as described may also provide regions, which are regarded as belonging to the object due to their gray levels, but in fact do not belong to it, like shadows or reflections. We get rid of them by simply choosing this segment as object, which is the closest to the center of the image (see Fig. 2c), centered segmentation.

Gabor transform and similarity function: The original image is convolved with a family of Gabor kernels, which differ in wavelength and orientation. We chose four wavelengths and eight orientations. The kernels take the form of a plane wave restricted by a Gaussian envelope function. At each image coordinate, we obtain filter responses for each Gabor wavelet. Filter responses at one image coordinate form a *jet J*. We can express the *i*th component of a jet in terms of amplitude a_i and phase ϕ_i :

$$J_i = (a_i, \phi_i). \tag{1}$$

Thus, a similarity function S between two jets J and J' can be defined as

$$S(J, J') = \frac{1}{2}\left(\frac{\sum_i a_i a_i' \cos(\phi_i - \phi_i')}{\sqrt{\sum_i a_i^2 \sum_i a_i' 2}} + 1\right). \tag{2}$$

S is the similarity function we used for our simulations, for the matching as well as for the tracking procedure. Its properties are described in Ref. [13].

Grid graphs: Given the result from the centered segmentation, we mask the original image of size 128×128 pixels with it. We cover the whole masked image with a grid graph of 13×13 vertices. All vertices are deleted, which lie on the background or which lie on the object but are too close to the background. The reason for this is to prevent vertices from incorporating too much information of the background. Then each vertex is labeled with the jet, which

corresponds to the position of the vertex. For display purposes the remaining vertices are connected by a minimal spanning tree (see Fig. 2e and f).

2.2. Matching object features

Elastic graph matching is described in detail in Ref. [14]. It has proved to be highly competitive in a number of recognition experiments [15–17] including the contests in the FERET program, a competitive benchmarking test organized by ARPA/ARL [18]. Given a graph with vertices labeled with jets, the aim of matching this graph to an image is to find new vertex positions which optimize the similarity of the vertex labels to the features extracted at the new positions. In the first stage (global move), the rigid graph is shifted across the image. For each position, we calculate the total similarity of the new positioned graph to the original graph. The total similarity is just the average similarity taken for each vertex by using the similarity function S. The position, which provides the highest similarity, is the starting position for the second stage which permits small graph distortions. The vertices are shifted in small surroundings of their starting positions. After this local move, the optimal position of the graph is found at the position which provides the highest total similarity.

2.3. Tracking object features

The tracking procedure we use is described in Ref. [19] and based on an idea reported in Refs. [20,21]. Given a sequence of a moving object and the pixel position of a landmark of the object for frame n, the aim is to find the corresponding position of the landmark in frame $n+1$. A similarity function S' between two jets J and J' is defined, which differs slightly from S

$$S'(J,J',\vec{d}) := \frac{\sum_i a_i a_i' \cos(\phi_i - \phi_i' - \vec{d}\vec{k}_i)}{\sqrt{\sum_i a_i^2 \sum_i a_i' 2}} \tag{3}$$

with $\vec{d}$ being the displacement vector of the two jets and $\vec{k}_i$ being the wave vectors of the Gabor filters. If J and J' are extracted at same pixel positions in the frames n and $n+1$, $\vec{d}$ (and thus the new position of the landmark) can be found by maximizing S' in its Taylor expansion. For each vertex of the graph of frame n, the displacements are calculated for frame $n+1$. Then a graph is created with its vertices at the new corresponding positions in frame $n+1$, and the labels of the new vertices are extracted from the new positions. To compensate for a subpixel error $\Delta\vec{d}$, the phases of the Gabor filter responses are shifted according to

$$\Delta\phi_i = \Delta\vec{d}\cdot\vec{k}_i. \tag{4}$$

2.4. Generation of view bubbles

For each view (i,j) of the hemisphere, an affiliated view bubble is created with the view as its center. To determine the view bubble, we compare neighboring views of (i,j) in all directions (east, west, north, and south). We match (or track) the grid graph of view (i,j) onto the neighboring views. If the similarities to the graph of view (i,j) are still sufficiently high, we depart one step further from (i,j). We begin with the views $(i-1,j)$ and $(i+1,j)$ (taking the continuity from the views $(0,j)$ to $(99,j)$ into account). If both views provide a sufficiently high similarity to the start graph, we go on with the views $(i-2,j)$ and $(i+2,j)$. We stop this procedure if one of both the tested views becomes too dissimilar. By doing the same for the vertical direction, we obtain four views $(i-n,j)$, $(i+n,j)$, $(i,j-m)$, and $(i,j+m)$, which define the view bubble for view (i,j). To depict it we draw an ellipse through these four views with view (i,j) as its center. Fig. 1 shows two ellipses projected onto the viewing hemisphere. For both the procedures, we used the same similarity threshold 0.77. A difference between tracking and matching concerning the similarity of views lies in the fact that during matching the similarity is calculated always with reference to the starting view, whereas during tracking the similarity refers to the preceding view.

3. Methods of comparison

As our work applies to arbitrary, real-world objects in contrast to artificially created CAD-objects and as the acquisition of object views is a very difficult and time-consuming process, the results of this work have been derived from the only two test objects. Thus, the properties of the chosen objects should provide challenging problems to the perception system and cover a wide range of possible object properties. The objects used in our experiments have been chosen with this in mind and they differ in the degree of their complexity. We chose a simple object (the 'Dwarf') and a more complex one (the cat 'Tom') (see Fig. 1). 'Simple' means that the views of the object do not change rapidly, while the object rotates. Such objects are often difficult to deal with in perception tasks such as pose estimation. The Dwarf is a relatively convex object with a rather similar shape for all viewing directions, whereas Tom is a more irregular object with faster changing views. Both objects vary in the degree of self-occlusions, which occur earlier for the Dwarf because of its sphere-like shape.

We used two different methods to compare the view bubbles generated by the matching (resp. tracking) procedure. With descriptive statistics, we made a quantitative comparison, and by judging the correspondences, which were found by both procedures, we compared both procedures qualitatively.

For the quantitative statistics, we determined the area of each view bubble by calculating the area of the ellipse, described in Section 2.4. (In Ref. [22], we also counted for each view, in how many other view bubbles it is

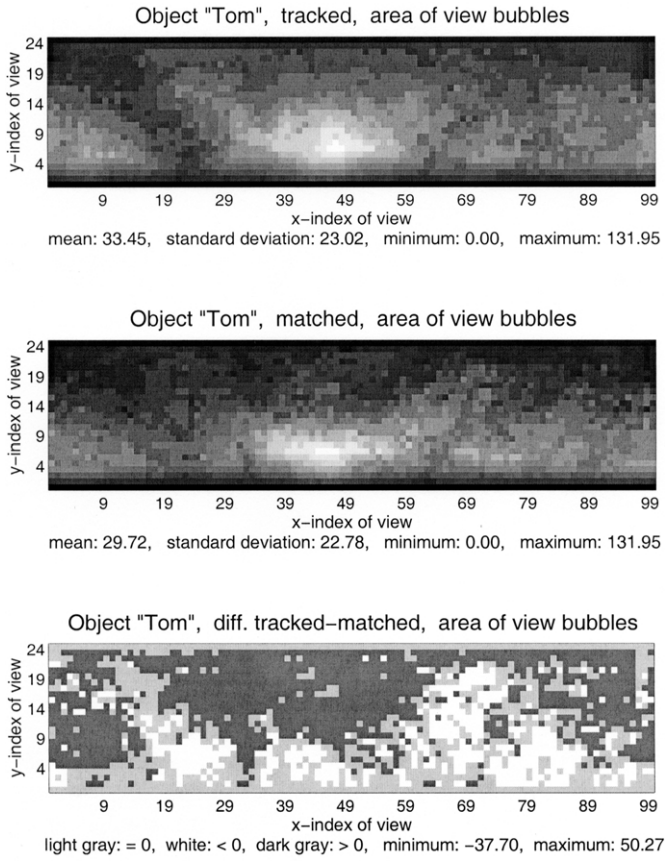

Fig. 3. Distribution of view similarities for object Tom.

contained. The results for this condition, however, resemble the results published here.) For both objects, we carried out a *t*-test to prove the hypothesis of different means of the areas of view bubbles for the tracked versus the matched view bubble samples.

For a qualitative comparison, we chose several sequences on the hemisphere (from a starting view to a destination view) for both objects, and the results for two of them are shown in Section 4. For each sequence, we performed the matching and the tracking procedure. To assess the correspondences by visible inspection, we displayed for each view of the sequences the resulting matched and tracked graphs and plotted the calculated similarities in diagrams. The sequences had an average size of about eight frames which means a covered rotation angle of 25.2°.

4. Results

The diagrams in Figs. 3 and 4 show results from the quantitative comparisons. Fig. 3 shows the distributions of areas of view bubbles for the object Tom, Fig. 4 for object Dwarf. For both figures, the first diagram depicts the results from the tracking procedure, the second the ones from the matching procedure. Lighter colors encode larger areas of view bubbles. Comparing the results for tracking and matching, the third diagram shows the difference between the first and second diagram.

From the diagrams, we get the following results. For both objects, Tom and Dwarf, the distribution of areas of view bubbles is qualitatively similar for the tracking procedure and for the matching procedure. The back view seen from slightly above and the front view provide the largest bubbles, and they can be regarded as canonical views.

These results hold for both objects, Tom and Dwarf. But there is a difference between the objects. The tracking procedure provides larger view bubbles than the matching procedure for the majority of views for the more complex object Tom, whereas for the more simple object Dwarf, it is the other way around: here the matching procedure provides larger view bubbles than the tracking procedure for the majority of views. The one-tailed *t*-test, with which we compared the mean values, was significant with $\alpha = 1\%$ for each case.

Figs. 5 (object Tom) and 6 (object Dwarf) show results for the qualitative comparison, i.e. the assessment of the correspondences. The first part of both figures displays views with graphs resulting from tracking (first row) and matching (second row). Both rows start with the starting view of the sequence. The next two images show the views where matching provided the last successfully matched and first mismatched graph in the sequence.

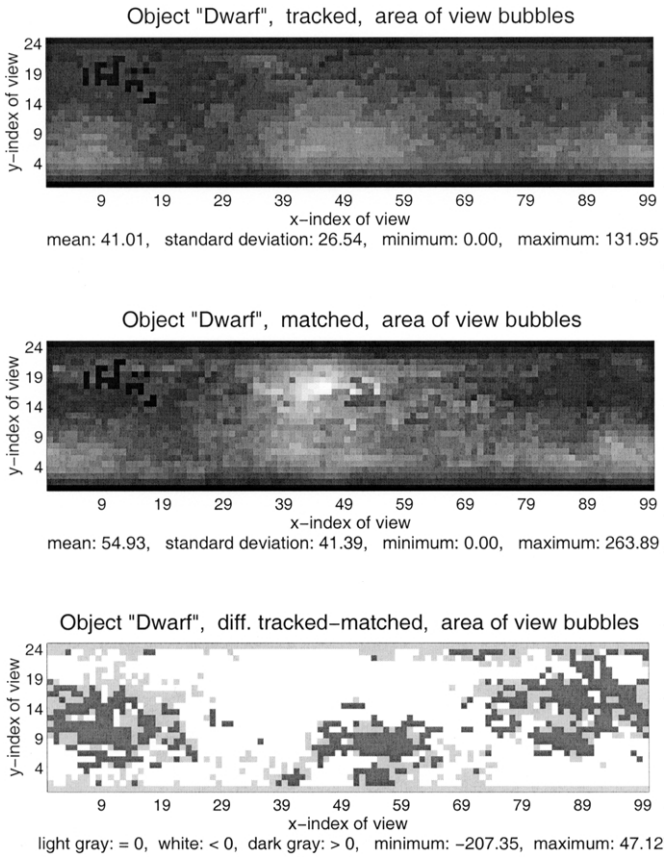

Fig. 4. Distribution of view similarities for object Dwarf.

Arrows point to mismatched vertices. The last images of the rows show the last views of the sequence where tracked graphs still keep the corresponding points, whereas the matched graphs do not. In the headers of the images the indices of the views are printed. ('Tr' means tracked, 'Mt' means matched.) For each view of both sequences, tracking provides the same or better correspondences than matching. The second part of both figures shows a diagram where the similarities for each view of the sequence to the starting view is plotted for tracking as well as matching. From the similarity diagrams of these and other analyzed sequences reported in Ref. [22], we get the following result. At the beginning of a sequence, the tracking procedure always provides higher similarities than the matching procedure. This relationship is reversed at that point of the sequence where the matching starts to provide poor correspondences, whereas tracking provides good correspondences until the end of the sequence (see Fig. 7).

5. Discussion and conclusion

Both procedures, matching and tracking of object features, are suitable to assess the pose robustness of views of a three-dimensional object by generating a distribution of view similarities on the viewing hemisphere. On the hemisphere areas of large and of small view bubbles arise. Centers of areas of large view bubbles can be regarded as canonical views (see Fig. 8) as they have been introduced in Section 1. This distribution of view similarities can guide the choice of views constituting a sparse, viewpoint-invariant object representation.

From both test objects, no statement on the superiority of one procedure in terms of size of view bubbles was possible because for the more complex object Tom tracking provided larger view bubbles, whereas matching outperformed tracking for the simpler object Dwarf. A possible explanation for this result could be that the rapidly changing views of object Tom cannot be matched over larger distances because the matching procedure is looking for the *same* appearance of the object features. We assume that the tracking procedure leads to larger view bubbles for 'complex' objects, whereas matching is superior for simple objects. Although we chose our test objects with respect to generality, this hypothesis has to be verified for more examples.

A reason for the more precise correspondences found by tracking could be the fact that an object feature changes its appearance, while the object rotates. The feature in the tracking procedure adapts to this change, whereas the matching procedure always searches for the same starting feature. The more the rotation proceeds the more difficult it is for the matching procedure to find the correct point,

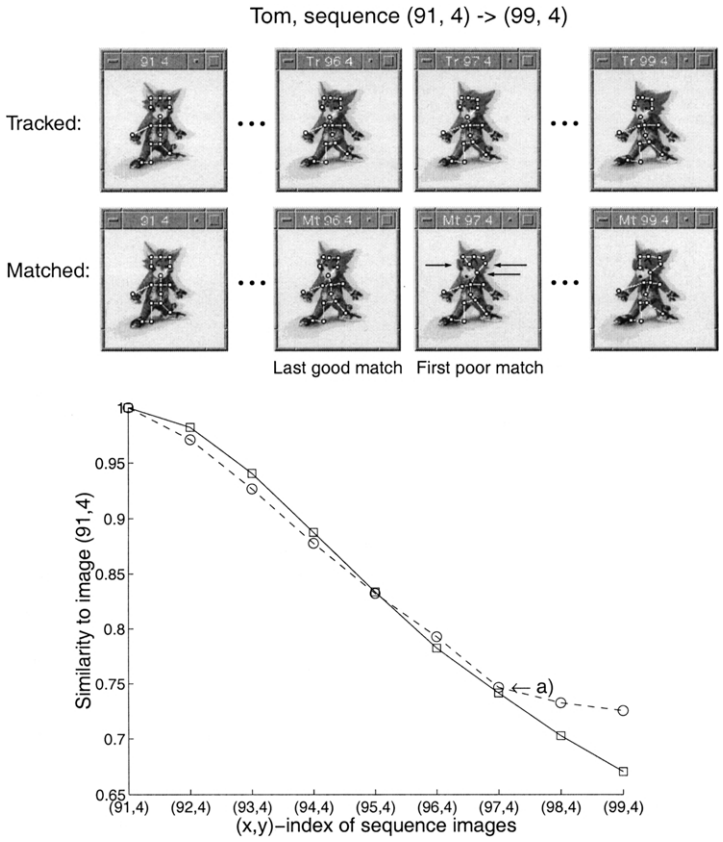

Fig. 5. Correspondences for object Tom: solid line represents tracked sequence, and dashed line represents matched sequence. (a) First poor match.

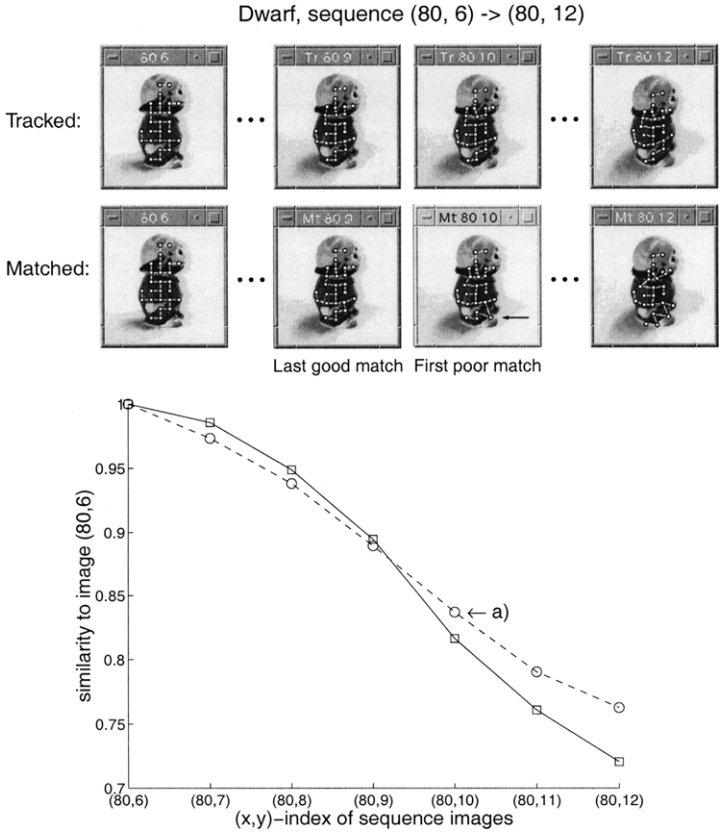

Fig. 6. Correspondences for object Dwarf: solid line represents tracked sequence, and dashed line represents matched sequence. (a) First poor match.

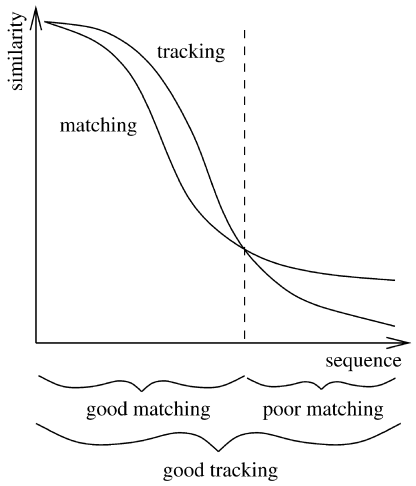

Fig. 7. Qualitative similarity diagram: 'good' and 'poor' is meant in the sense of correct, respectively, incorrect, correspondences. See description in the text for details.

whereas the tracking utilizes *continuous information*. Matching is a more appropriate method if the task is to find features with the *same* appearance, and tracking is more appropriate if *changes* of the features should be followed.

In other words, good correspondences are derived from the *continuity* of successive views and not from *disconnected* static views. This result has a number of parallels to primate object perception. The experiments of Kellman [23] with infants suggest that they have the ability to perceive the three-dimensional form of an object only if information about continuous optical transformations given by motion is available. They are not able to apprehend the overall form of an object from static views, even if they are multiple or sequential. Another result is furnished by Harman and Humphrey [24]. They claim that different object representations are generated, depending on the presentation of either regular or random sequences of views of the object. When a sequence of rotation is encoded, the associated temporal context may lead to the construction of a linked, higher-order system of representations for a given object, whereas without temporal context, a single representation of each object view may be constructed. Also some physiological reasons emphasize the importance of the successive appearance of views for the learning of an object representation. Miyashita [25] trained monkeys to match complex fractal patterns, which were presented successively in a fixed series of 100 items. After training, some cells in the anterior temporal cortex were found to show selectivity for a small number of patterns which had been presented successively. This gives evidence for learning based on temporal associations rather than on pattern overlap. Perrett et al. [26] already observed that an object representation in the form of a collection of stored views is structured in the sense that views belonging together because of their successive appearance are more closely associated with each other in the representation. This was confirmed in a later study by Edelman and Weinshall [27].

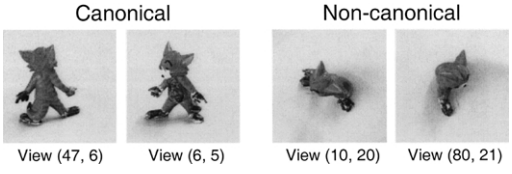

Fig. 8. Canonical and non-canonical views for object Tom. View (47,6) is the view with the largest area of its view bubble (generated by tracking). Compare with the first diagram of Fig. 3.

Even if it would turn out that matching is superior to tracking for simple objects in terms of size of the view bubbles, we suggest that precise correspondences should take priority over larger view bubbles, particularly for further processing. For a view interpolation, e.g. precise correspondences are necessary, and to establish these correspondences, the continuity information of successive views has to be utilized. Accordingly, our final conclusion is that tracking of object features is superior to matching for measuring the pose robustness of object views, especially for complex objects.

Acknowledgements

Many thanks are due to Pervez Mirza for proofreading this article.

References

[1] A.R.J. Francois, G.G. Medioni, Interactive 3D model extraction from a single image, Image and Vision Computing 19 (6) (2001) 317–328.
[2] H. Murase, S. Nayar, Visual learning and recognition of 3-D objects from appearance, International Journal of Computer Vision 14 (1995) 5–24.
[3] J.B. Tenenbaum, V. de Silva, J.C. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2322.
[4] H.S. Seung, D.D. Lee, The manifold ways of perception, Science 290 (2000) 2268–2269.
[5] V. Blanz, B. Schölkopf, H.H. Bülthoff, C. Burges, V. Vapnik, T. Vetter, Comparison of view-based object recognition algorithms using realistic 3D models, in: C.v.d. Malsburg, W.v. Seelen, J.C. Vorbrüggen, B. Sendhoff (Eds.), Artificial Neural Networks—ICANN'96, Berlin, Springer Lecture Notes in Computer Science, vol. 1112, 1996, pp. 251–256.
[6] B. Schölkopf, Support Vector Learning, PhD Thesis, Informatik der Technischen Universität Berlin, 1997.
[7] S.E. Palmer, E. Rosch, P. Chase, Canonical perspective and the perception of objects, in: I. Long, A. Baddeley (Eds.), Attention and Performance IX, Erlbaum, Hillsdale, NJ, 1981, pp. 135–151.
[8] R. Malik, T. Whangbo, Angle densities and recognition of 3D objects, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (1) (1997) 52–57.
[9] D. Weinshall, M. Werman, On view likelihood and stability, IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (2) (1997) 97–108.
[10] J.J. Koenderink, A.J.v. Doorn, The internal representation of solid

shape with respect to vision, Biological Cybernetics 32 (1979) 211–216.
[11] M. Seibert, A.M. Waxman, Adaptive 3-D object recognition from multiple views, IEEE Transactions on Pattern Analysis and Machine Intelligence 14 (2) (1992) 107–124.
[12] J.C. Vorbrüggen, Zwei Modelle zur datengetriebenen Segmentierung visueller Daten, Reihe Physik, vol. 47, Verlag Harri Deutsch, Thun, Frankfurt am Maim, 1995.
[13] L. Wiskott, Labeled Graphs and Dynamic Link Matching for Face Recognition and Scene Analysis, Reihe Physik, vol. 53, Verlag Harri Deutsch, Thun, Frankfurt am Main, Germany, 1995 PhD Thesis.
[14] M. Lades, J.C. Vorbrüggen, J. Buhmann, J. Lange, C.v.d. Malsburg, R.P. Würtz, W. Konen, Distortion invariant object recognition in the dynamic link architecture, IEEE Transactions on Computers 43 (1993) 300–311.
[15] L. Wiskott, J.-M. Fellous, N. Krüger, C.v.d. Malsburg, Face recognition by elastic bunch graph matching, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 775–779.
[16] R.P. Würtz, Object recognition robust under translations, deformations and changes in background, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 769–775.
[17] M. Becker, E. Kefalea, E. Maël, C.v.d. Malsburg, M. Pagel, J. Triesch, J.C. Vorbrüggen, R.P. Würtz, S. Zadel, GripSee: a gesture-controlled robot for object perception and manipulation, Autonomous Robots 6 (2) (1999) 203–221.
[18] K. Okada, J. Steffens, T. Maurer, H. Hong, E. Elagin, H. Neven, C.v.d. Malsburg, The Bochum/USC face recognition system and how it fared in the FERET phase III test, in: H. Wechsler, P.J. Phillips, V. Bruce, F. Fogeman Soulié, T.S. Huang (Eds.), Face Recognition: From Theory to Applications, Springer, Berlin, 1998, pp. 186–205.
[19] T. Maurer, C.v.d. Malsburg, Tracking and learning graphs and pose on image sequences of faces, Proceedings of the Second International Conference on Automatic Face- and Gesture-Recognition, Killington, Vermont, USA, October 1996, pp. 176–181.
[20] D.J. Gleet, A.D. Jepson, Computation of component image velocity from local phase information, International Journal of Computer Vision 5 (1) (1990) 77.
[21] W.M. Theimer, H.A. Mallot, Phase-based binocular vergence control and depth reconstruction using active vision, CVGIP: Image Understanding 60 (3) (1994) 343.
[22] G. Peters, B. Zitova, C.v.d. Malsburg, A comparative evaluation of matching and tracking object features for the purpose of estimating similar-view-areas of 3-dimensional objects, Internal Report IRINI 99-06, Institut für Neuroinformatik, Ruhr-Universität Bochum, D-44780 Bochum, Germany, April 1999.
[23] P.J. Kellman, Perception of three-dimensional form in infancy, Perception and Psychophysics 36 (1984) 353–358.
[24] K.L. Harman, G.K. Humphrey, Encoding 'regular' and 'random' sequences of views of novel 3D objects rotating in depth, Investigative Ophthalmology and Visual Science, ARVO Annual Meeting, May 10–15 1998, vol. 39(4), Fort Lauderdale, Florida, USA, 1998 p. 856.
[25] Y. Miyashita, Associative representation of visual long term memory in the neurons of the primate temporal cortex, in: E. Iwai, M. Mishkin (Eds.), Vision, Memory and the Temporal Lobe, Elsevier, New York, 1990, pp. 75–87.
[26] D.I. Perrett, A.J. Mistlin, A.J. Chitty, Visual neurons responsive to faces, Trends in Neurosciences 10 (1989) 358–364.
[27] S. Edelman, D. Weinshall, A selforganizing multiple-view representation of 3D objects, Biological Cybernetics 64 (12) (1991) 209–219.