ROBUST OBJECT TRACKING BASED ON SPATIAL CHARACTERIZATION OF OBJECTS BY ADDITIVE INVARIANTS

Holger Eggers[1] *Fabrice Moscheni*[2] *Roberto Castagno*[2]

[1]Technische Informatik I, Technische Universität Hamburg–Harburg
D-21071 Hamburg, Germany
[2]Signal Processing Laboratory, Swiss Federal Institute of Technology
CH-1015 Lausanne, Switzerland

ABSTRACT

We present an improved object tracking algorithm in the context of spatio-temporal segmentation. By incorporating invariants for the spatial characterization, the information supplied by the tracking algorithm to the current segmentation is extended from a purely temporal to a more comprehensive spatio–temporal description of the objects in the scene. Thereby, the extraction and the tracking of meaningful objects in video sequences is enhanced. The proposed spatial characterization is shown to be efficiently implementable due to the additivity in feature space of the chosen class of invariants.

1. INTRODUCTION

This paper addresses the problem of automatically detecting and tracking arbitrary objects in video sequences. The spatio-temporal segmentation has proven to provide a reliable means of completely decomposing dynamic scenes into their constituent objects [1, 3]. Since it usually uses only the preceding and the current frame of a sequence to partition the latter, tracking has to be applied to assure the coherence and stability of the segmentation through time.

We propose in this work an improved tracking algorithm. It is derived from the approach described in [2, 4] which divides the tracking into two distinct steps. In the first step, information about the objects in previous frames is gathered and supplied to the current segmentation process. In the second step, the objects in the current frame are linked to the corresponding ones in previous frames.

We focus here on improving the first step. The spatio–temporal segmentation employed for the decomposition of the sequence is based on a region merging approach. The regions of an initial partition of the current frame are iteratively combined to meaningful groups. These groups are then regarded as objects [1, 3]. Two regions are likely to be merged into the same object if their spatio-temporal similarity is high. The proposed measure of similarity considers information from previous frames to stabilize the partition of the sequence through time.

To measure this similarity, the tracking algorithm presented in [2, 4] exploits temporal information only in the form of a prediction of each object's position in the current frame. Thus, the region merging becomes unreliable due to the generally error–prone motion estimation.

This work was carried out while the first author was a visiting student at the Signal Processing Laboratory of the Swiss Federal Institute of Technology in Lausanne. It was partially supported by the European long term research project ESPRIT 20229 'NOBLESSE'.

In this paper, we propose a similarity measure based on the characterization of objects by invariant features. This method allows us to exploit spatial information relative to previous frames while segmenting the current one. The calculation of this similarity measure is divided into two successive steps.

In the first step, a searching algorithm determines the combinations of regions in the current frame which best match the objects in previous frames with respect to their invariants.

In the second step, pairs of regions which are contained in the same combination are attributed a high similarity, whereas pairs whose elements belong to different combinations are set to lower similarity. This mechanism proves more robust than simply directly merging all regions that are grouped in a well matched combination. Besides, this similarity measure can be integrated into the tracking algorithm, in order to combine spatial and temporal information from previous frames. Finally, this approach is less sensitive to noisy measurements of the invariants, as shown in the example in Section 5.

This paper is structured as follows. The selected class of invariants and related properties are summarized in Section 2. Section 3 presents the search algorithm, and Section 4 describes the computation of the similarity measure. Experimental results are given in Section 5. Eventually, Section 6 draws some conclusions.

2. INVARIANT FEATURES

Invariant features provide a spatial characterization of objects which remains unchanged under the action of certain transformation groups. In [6], a class of invariants for grey scale images is proposed which allows, in contrast to the predominantly used moment invariants such as [9], an efficient computation of the features of combinations of regions. It is constructed by averaging, over the transformation group of rotations and translations, polynomial functions of the grey values of images. Features of individual objects are extracted by restricting the area for which the invariants are computed to the part of the image that is covered by the respective object [8].

The polynomials proposed in [7] are used to construct a set of twelve invariants forming a feature vector for the spatial characterization of objects. The low order of the polynomials assures its robustness with respect to noise. Since a significant deviation of just one invariant already indicates a major difference in shape or texture between two objects, we use the infinite norm for comparisons of feature vectors.

On a spatially discrete grid, the averaging over the considered transformation group has to be approximated. A sum

of interpolated values replaces the analytical integration over the rotation angle and the translation vector. Experimental investigations suggest that the bilinear interpolation leads, despite its low-pass characteristic, to superior results in comparison with other simple interpolation techniques such as the nearest neighbor interpolation. In order to assure uniform accuracy of the features values, the number of summands considered by the approximation should be adapted to the support size [7] of the employed polynomials.

An invariant feature is said to be additive in the feature space if its value for the union of two regions is equal to the sum of its values for the two individual regions. Exact additivity is reached for the invariants only when the regions are separated by more than the maximal support size of the used polynomials (i.e. the set of pixels on which the polynomials is evaluated don't overlap). In a segmented image, however, neighboring regions are not sufficiently separated. Therefore, only approximated additivity is achieved. The resulting error accumulates when several regions are combined into one object and thus has to be corrected. This is achieved by introducing correction terms [5]. They can be simultaneously calculated with the invariant features of the individual regions and thereby cause no significant computational overhead.

The invariance of the introduced features is restricted to rotational and translational transformations. Since usually an affine motion model given by

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = \begin{pmatrix} a_2 & a_3 \\ a_5 & a_6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + \begin{pmatrix} a_1 \\ a_4 \end{pmatrix}$$

is used to describe the movement of objects, the features of objects will only remain unchanged if the affine parameters satisfy

$$\begin{aligned} \begin{vmatrix} a_2 & a_3 \\ a_5 & a_6 \end{vmatrix} &= 1 \\ a_2 - a_6 &= 0 \\ a_3 + a_5 &= 0 \end{aligned}$$

In practice, these conditions are generally not met if a movement of the camera occurs. Therefore, the global motion compensation is first applied. Furthermore, the robustness of the feature vectors with respect to zoom is improved by normalizing the invariants for the comparison of spatial characteristics with the area of the corresponding objects. The restriction of the invariance to rotations and translations therefore constitutes no substantial limitation to the applicability of these features.

3. SEARCHING ALGORITHM

The searching algorithm described in this section tests all combinations of regions for their spatial similarity with the reference objects in the previous frame. The theoretical upper limit for the number of combinations that have to be tested is given by

$$\sum_{i=1}^{n} \binom{n}{i} = 2^n - 1$$

where n denotes the number of regions in the initial segmentation. It corresponds to an exponential computational complexity $\mathcal{O}(2^n)$. Even for the rather small values of n that usually occur in video sequences, a considerable reduction of the required computation time is necessary. Hence, the number of tested combinations must be significantly reduced and the complexity of the calculation of the invariants for each combination must be minimized.

The first goal is achieved by only considering neighboring regions as elements of a combination. For this purpose, the initial segmentation of the frame is represented as a graph. All allowed combinations on this graph are successively generated by an iterative algorithm based on a queue [5]. Their invariant features are compared with those of the reference objects and the respective best matches are stored in lists. Since the graph usually contains cycles, some combinations are repeatedly generated. Therefore, we test whether a generated combination is already included in the queue or not. The complexity of this test essentially influences the computation time of the whole algorithm. For an efficient implementation we propose to use a hash table which keeps track of the already considered combinations. In order to access this hash table, the combinations are represented by a binary code C. Simulations have confirmed that the distribution of C is adequately smoothed by a simple hash function H such as

$$H(C) = C \bmod p$$

where p denotes a prime number chosen dependent on the expected overall number of combinations.

The second goal is attained by exploiting the property of additivity in feature space of the chosen class of invariants. The complete recalculation of the features of each combination would results in an unacceptable computation time. Therefore, the invariants of the separate regions and the correction terms are calculated once for all prior to the start of the search. The features of combinations of regions are then derived according to

$$\begin{aligned} \vec{I}_{M_1} &= \vec{I}_{m_1} \\ \vec{I}_{M_{k+1}} &= \vec{I}_{M_k} + \vec{I}_{m_{k+1}} + \sum_{i=1}^{k} \vec{I}_{m_i, m_{k+1}} \\ &\quad + \sum_{\substack{i,j=1 \\ i \neq j}}^{k} \vec{I}_{m_i, m_j, m_{k+1}} + \dots \end{aligned}$$

where $\vec{I}_{M_k}$ denotes the invariant vector for the current combination of k regions, $\vec{I}_{m_k}$ that for region k and $\vec{I}_{m_i, m_j}, \vec{I}_{m_i, m_j, m_k}, \dots$ the correction terms.

4. SIMILARITY MEASURE

When combinations of regions are matched against the reference objects in the previous frame, the best match does not stand out unambiguously with regard to the other combinations (see Table 1 and Sec. 5). Therefore, the result of this matching is not robust enough for an appropriate segmentation and the subsequent tracking phase. It is therefore necessary to exploit the information provided not only by the best, but by a group of most likely combinations. The basic underlying idea is that a pair of regions that appear together in most of the top ranked combinations can be assigned a very high similarity index, which will very likely result in their merging.

We propose the following approach to compute the similarity measure. The pairs of regions are searched in the

lists of best matching combinations. If many well–matching combinations contain both of them, their similarity measure is increased. If, on the other hand, the regions are often split in different combinations, the degree of similarity is decreased.

In order to consider all occurrences of the two regions and to quantify the similarity measure, we assign suitable weights to the individual combinations. For this purpose, the distribution of the normalized invariant features through time is analyzed for each object. We model it by a Gaussian distribution since the causes of the deviations of the features are diverse and hard to describe mathematically. Simulations confirmed the sufficient accuracy of this model [5].

The normalization of the invariants renders the comparison of spatial characteristics more robust. However, this introduces ambiguities because some combinations are only distinguishable by their area. Therefore, the weight assigned to a combination should take information about the area of the regions into account as well.

We propose to weight each combination t with

$$w_t = \frac{1}{2\pi\sigma_A\sigma_I} \exp\left(-\frac{(t_A-\mu_A)^2}{2\sigma_A^2} - \frac{(t_I-\mu_I)^2}{2\sigma_I^2}\right), \quad (1)$$

where t_A and t_I denote the area and a normalized invariant of t, and μ_A, μ_I, σ_A^2, σ_I^2 the mean and the variance of the corresponding distributions, respectively. The selected invariant is the one which yields the maximal relative error. The distributions are based on all detected occurrences of the respective object in previous frames.

The similarity measure for regions i and j is then defined as

$$S_{ij} = \frac{\sum\limits_{\substack{t_k \\ i,j\in t_k}} w_{t_k} - \sum\limits_{\substack{t_k \\ i\in t_k,\, j\notin t_k}} w_{t_k} - \sum\limits_{\substack{t_k \\ i\notin t_k,\, j\in t_k}} w_{t_k}}{\sum\limits_{\substack{t_k \\ i,j\in t_k}} w_{t_k} + \sum\limits_{\substack{t_k \\ i\in t_k,\, j\notin t_k}} w_{t_k} + \sum\limits_{\substack{t_k \\ i\notin t_k,\, j\in t_k}} w_{t_k}}, \quad (2)$$

where t_k represent the combinations of regions.

5. EXPERIMENTAL RESULTS

In this section, experimental results are presented. The investigated example is taken from the sequence "Table Tennis".

Figure 1. Final spatio–temporal segmentation of the preceding frame which contains 4 reference objects

Figure 1 depicts the final spatio–temporal segmentation of the preceding frame in which four reference objects have been defined. These are the background, the bat and hand, the arm, and the ball.

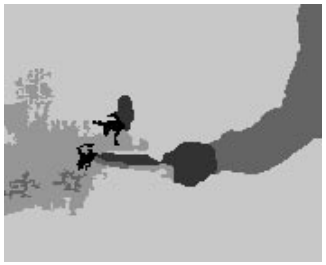

Figure 2. Initial segmentation of the current frame which contains 13 regions

The initial segmentation of the current frame yields thirteen regions as shown in Figure 2. Eight of them correspond to the background and three to the bat and the hand.

In Table 1, the combinations of regions in the current frame yielding the five best matches for each object are listed. They are determined by the described searching algorithm. The thirteen regions are displayed in this table according to their size and the object to which they actually correspond. The combinations are ordered according to the maximal relative deviation of their cumulative invariants from the invariants of the corresponding object. A '1' indicates that the region is contained in the considered combination, a '0' that it is not.

	Relative error	Combination of regions			
		Back–ground	Bat, hand	Arm	Ball
Back–ground	0.00182773	11111111	000	0	0
	0.00209014	11111110	000	0	0
	0.00361198	11111111	001	0	0
	0.00365904	11111110	001	0	0
	0.00377789	11101111	000	0	0
Bat, Hand	0.01324284	00000000	111	0	0
	0.03807617	00000100	111	0	0
	0.04794542	00000100	110	0	0
	0.07185463	00000000	110	0	0
	0.08567622	00000100	101	0	0
Arm	0.00480967	00000000	000	1	0
	0.69369870	00000000	001	0	0
	0.97235459	00000000	011	0	0
	1.13118839	00000000	010	0	0
	2.19361901	01000001	011	0	0
Ball	0.25495994	00000000	000	0	1
	0.50857323	00010000	000	0	1
	0.68332916	00000000	100	0	0
	0.69472170	00000100	100	0	0
	0.70333320	00000000	101	0	0

Table 1. Lists of best matching combinations of regions obtained by the proposed searching algorithm for each of the four reference objects

For all four objects, the best match exactly comprises all those regions which actually correspond to the respective object. The table also shows that the second best match does not always differ from the best significantly in terms of relative error. On the other hand, it is possible to see at first glance that the pairs of regions to be merged can be identified in a fairly reliable and robust way by taking into account not only the best matches, but for example the 5 best ones, as described in Sec. 4.

	Background								Bat, Hand			Arm	Ball
Back–ground	1.00	1.00	1.00	0.99	1.00	0.14	1.00	0.24	-1.00	-1.00	0.28	-1.00	-1.00
	1.00	1.00	1.00	0.99	1.00	0.14	1.00	0.24	-1.00	-1.00	0.28	-1.00	-1.00
	1.00	1.00	1.00	0.99	1.00	0.14	1.00	0.24	-1.00	-1.00	0.28	-1.00	-1.00
	0.99	0.99	0.99	1.00	0.99	0.14	0.99	0.24	-1.00	-1.00	0.28	-1.00	-1.00
	1.00	1.00	1.00	0.99	1.00	0.14	1.00	0.24	-1.00	-1.00	0.28	-1.00	-1.00
	0.14	0.14	0.14	0.14	0.14	1.00	0.14	-0.29	-0.25	-0.25	-0.13	-1.00	-1.00
	1.00	1.00	1.00	0.99	1.00	0.14	1.00	0.24	-1.00	-1.00	0.28	-1.00	-1.00
	0.24	0.24	0.24	0.24	0.24	-0.29	0.24	1.00	-1.00	-1.00	-0.23	-1.00	-1.00
Bat, Hand	-1.00	-1.00	-1.00	-1.00	-1.00	-0.25	-1.00	-1.00	1.00	0.99	-0.68	-1.00	-1.00
	-1.00	-1.00	-1.00	-1.00	-1.00	-0.25	-1.00	-1.00	1.00	0.99	-0.68	-1.00	-1.00
	0.28	0.28	0.28	0.28	0.28	-0.13	0.28	-0.23	-0.68	-0.68	1.00	-1.00	-1.00
Arm	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1.00	-1.00
Ball	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	-1.00	1.00

Table 2. Values of the similarity measure for pairs of regions for the given example

In Table 2, the results of the evaluation of the proposed similarity measure for the given example are listed. Ideally, the similarity of two regions which both actually correspond to the same object should be 1.0 and that of all other pairs -1.0. This is attained for the arm and the ball. In practice, however, positive values instead of 1.0 and negative values instead of -1.0 are usually sufficient. Except for the smallest region of the bat and the hand, this is achieved for all objects.

6. CONCLUSION

A new similarity measure based on invariant features has been proposed which increases the robustness of video sequence segmentation. The proposed measure has been introduced in order to evaluate the probability that pairs of regions, resulting from a segmentation of a frame in a video sequence, actually belong to the same object and can reasonably be merged into a single region. The measure of similarity depends on the result of a number of tests in which the invariant features of groups of regions in the current frame are matched with the features of objects detected in the previous frames. In order to reduce the computational complexity, the method exploits the property of additivity in feature space of the proposed invariant features, which therefore need to be calculated only once and then added up to obtain the global values corresponding to combinations of regions.

The results show that this similarity measure efficiently allows to propagate spatial characteristics of objects through time. Thereby, the information supplied by the tracking algorithm to the current segmentation is enriched, and the resulting decomposition of video sequences is more coherent.

REFERENCES

[1] F. Moscheni, F. Dufaux, "Region Merging based on Robust Statistical Testing", *SPIE Proceedings of the VCIP*, Orlando, U.S.A., March 1996

[2] F. Moscheni, F. Dufaux, M. Kunt, "Object Tracking based on Temporal and Spatial Information", *IEEE Proceedings of the ICASSP*, Vol. 4, pp. 1914–1917, Atlanta, U.S.A., May 1996

[3] F. Moscheni, S. Bhattacharjee, "Robust Region Merging for Spatio–Temporal Segmentation", *IEEE Proceedings of the ICIP*, Vol. 1, pp. 501–504, Lausanne, Switzerland, October 1996

[4] F. Moscheni, F. Ziliani, "Object Tracking based on Temporal and Spatial Information", *ISO/IEC JTC1/SC29/WG11 MPEG96/M0962*, 1996

[5] H. Eggers, *Analysis and Improvement of a Video Sequence Segmentation and Object Tracking Tool*, Diploma thesis, Swiss Federal Institute of Technology, Lausanne, Switzerland, 1996

[6] H. Schulz–Mirbach, *Anwendung von Invarianzprinzipien zur Merkmalsgewinnung in der Mustererkennung*, Ph.D. thesis, Technische Universität Hamburg-Harburg, Germany, Fortschrittberichte VDI, Reihe 10, Nr. 372, VDI–Verlag, 1995

[7] H. Schulz–Mirbach, "Invariant Features for gray scale Images", in G. Sagerer, S.Posch, F. Kummert, *17. DAGM Symposium Mustererkennung*, pp. 1–14, Springer, 1995

[8] H. Schulz–Mirbach, H. Burkhardt, S. Siggelkow, "Using Invariant Features for Content based Data Retrieval", *Proceedings of the NOBLESSE Workshop*, Lausanne, Switzerland, September 1996

[9] G. Taubin, D. B. Cooper, "Object Recognition based on Moment (or Algebraic) Invariants", in J. L. Mundy, A. Zisserman, *Geometric Invariance in Computer Vision*, pp. 375–397, MIT Press, 1992