Detection of moving objects in video using a robust motion similarity measure

Hieu T. Nguyen, Marcel Worring, and Anuj Dev

Abstract— **This correspondence deals with the segmentation of a video clip into independently moving visual objects. This is an important step in structuring video data for storage in digital libraries. The method follows a bottom-up approach. The major contribution is a new well founded measure for motion similarity leading to a robust method for merging regions. The improvements with respect to existing methods have been confirmed by experimental results.**

Keywords— **motion similarity, motion segmentation, content-based indexing, video analysis**

I. Introduction

Video is becoming an important datatype in digital libraries. Besides the traditional verbal user queries, the library should support queries based on object shape and motion. Objects and their characteristics therefore form basic units for content-based retrieval and presentation of video. They are further useful for video compression and the creation of hypervideo documents.

Our correspondence deals with motion segmentation, i.e. the decomposition of a video scene into independently moving visual objects. Starting from an oversegmentation of the scene, it merges regions based on motion similarity.

The paper is structured as follows. In Section II we formulate the problem and provide a short review of existing literature. Our new region merging method is described in Section III. Section IV shows the results of applying the method on some standard test image sequences.

II. Problem formulation

Suppose there are K moving rigid objects in the scene, we want to recover their regions of projection $C_1, ..., C_K$ on the image plane. When a surface of a rigid moving object is planar or distant enough from the camera, its optic flow at position $\mathbf{x} = (x, y)$ is well modeled by the quadratic transformation [1]:

$$\begin{aligned} v_x &= a_1 + a_2 x + a_3 y + a_7 xy + a_8 x^2 \\ v_y &= a_4 + a_5 x + a_6 y + a_7 y^2 + a_8 xy \end{aligned} \quad (1)$$

where $\boldsymbol{\vartheta} = (a_1, \ldots, a_8)$ are the motion parameters of the moving surface. In existing literature the modelling of the optic flow of a region by a parametric model is used either in the direct form (1) or in combination with the intensity matching equation. In both cases we can consider it in the generalized form:

$$\mathbf{y}(\mathbf{x}) = \mathbf{f}(\mathbf{x}; \boldsymbol{\vartheta}_k) + \boldsymbol{\varepsilon}_\mathbf{x} \quad \forall \mathbf{x} \in C_k; \quad k = 1, ..., K \quad (2)$$

where $\mathbf{y}(\mathbf{x})$ is the measurement at pixel $\mathbf{x}$. In (1) $\mathbf{y}$ would correspond to the vector (v_x, v_y). $\mathbf{f}$ is a known function. $\boldsymbol{\vartheta}_k$ is the vector of motion parameters of region C_k which is unknown a priori and is to be estimated together with C_k. The term $\boldsymbol{\varepsilon}_\mathbf{x}$ represents the measurement error. The problem can then be viewed as segmenting the data, which are described by multiple regression models, into groups so that data in each group is described by a common model. The challenge is that both C_k and $\boldsymbol{\vartheta}_k$ are not known a priori.

First, a granularity level for regions has to be selected. Pixel-wise classification is inherently not reliable due to the aperture problem as pixels in homogeneous regions can be assigned to any model. Trying to overcome this problem recent methods start from homogeneous regions, usually obtained through intensity or color segmentation. Then regions are merged according to motion based criteria. Several criteria have been proposed [1], [8], [5], [2]. The earliest method [8] used a scaled Euclidean distance. This measure has been criticized to be sensitive to inaccuracy in parameter estimation as well as the choice of the scaling matrix [5], [2]. Furthermore, we have shown that the merging results depend on the choice of the origin of the coordinate system [6]. In [9] the merging decision is based on whether residuals in two regions can be expressed by one Gaussian distribution. We found that the measure tests for the difference in motion parameters of the regions as well as difference in noise variance. The latter test is, in fact, not needed. Altunbasak *et al* [2] proposed a merging procedure minimizing the residual over all regions. The global minimum corresponds to the maximum likelihood solution for both region labels and motion parameters. The problem is that the objective function usually has very many minima and the proposed technique finds a local minimum only. A good starting point is therefore required. An elaborate method was developed by Moscheni *et al* [5]. However, an asymmetric similarity measure is used. Moreover, the method depends on a large number of parameters that need to be set.

In conclusion, existing merging criteria are still ad-hoc. To this end, we propose a new rigorous approach for definition of motion similarity and develop a new merging method utilizing the strengths of the methods of [9] and [2] and at the same time overcoming their shortcomings.

III. New robust method for region merging.

A. Motion statistics based region similarity

Let $R_1, ..., R_M$ be the regions obtained from the initial oversegmentation. We assume that the initial segmentation is such that there are no regions occluding the object boundaries, hence, each R_i is a subregion of one C_k. Let $\boldsymbol{\theta}_i$ be the vector of motion parameters of R_i, then $\boldsymbol{\theta}_i \in \{\boldsymbol{\vartheta}_1, ..., \boldsymbol{\vartheta}_K\}$. Obviously, two regions undergo a common motion and therefore are supposed to be merged if and only if $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$. In practice, however, we have to base our decision on an estimate $\hat{\boldsymbol{\theta}}_i$ of $\boldsymbol{\theta}_i$ which is commonly

The authors are with the Intelligent Sensory Information Systems group, University of Amsterdam, Faculty of WINS, Kruislaan 403, NL-1098 SJ, Amsterdam, The Netherlands. Email: { tat,worring,anuj }@wins.uva.nl

obtained via least squares minimization:

$$\hat{\boldsymbol{\theta}}_i = \arg\min_{\boldsymbol{\theta}} S_i(\boldsymbol{\theta}) \quad \text{where} \quad S_i(\boldsymbol{\theta}) = \sum_{\mathbf{x}\in R_i} |\mathbf{y}(\mathbf{x}) - \mathbf{f}(\mathbf{x};\boldsymbol{\theta})|^2 \tag{3}$$

In the presented method we first apply the robust estimation technique given in [4] to detect outliers defined as pixels whose measurement is not described by the same model as the majority of pixels in the region. Then the parameters are reestimated for the clean data using least-square. Although the influence of the outliers is reduced, $\hat{\boldsymbol{\theta}}$ may be different from the true $\boldsymbol{\theta}$ due to noise. As a consequence, two regions of the same object may turn out to have different estimated motion parameters. Therefore, inference about the equality of $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$ should be made by means of a statistical test.

We need to test the hypothesis :

$$H_0 : \boldsymbol{\theta}_i = \boldsymbol{\theta}_j \quad \text{versus} \quad H_1 : \boldsymbol{\theta}_i \neq \boldsymbol{\theta}_j \tag{4}$$

The problem is much like testing whether two sets of data

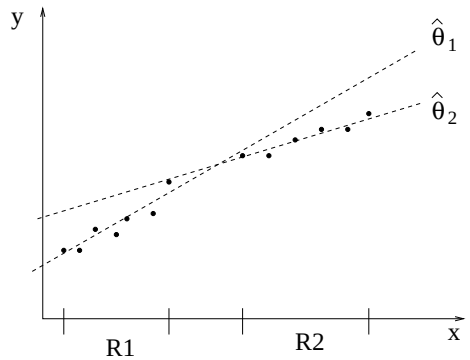

Fig. 1. *Test if two sets of data points represent the same model.*

points represent the same line (see Figure 1). Let L be the likelihood function of the measurements in R_i and R_j, given $\boldsymbol{\theta}_i$ and $\boldsymbol{\theta}_j$:

$$L(\mathbf{y}(\mathbf{x}), \mathbf{x} \in R_i \cup R_j | \boldsymbol{\theta}_i, \boldsymbol{\theta}_j) = (2\pi\sigma_i^2)^{-n_i/2}(2\pi\sigma_j^2)^{-n_j/2} \times \exp\{-[\frac{1}{2\sigma_i^2}S_i(\boldsymbol{\theta}_i) + \frac{1}{2\sigma_j^2}S_j(\boldsymbol{\theta}_j)]\} \tag{5}$$

where σ_i and n_i are the variance of the measurement errors and the number of pixels in R_i respectively. We construct the log-likelihood ratio statistic:

$$\Delta(i,j) = -2\log(\sup_{\boldsymbol{\theta}_i=\boldsymbol{\theta}_j} L / \sup_{\boldsymbol{\theta}_i\neq\boldsymbol{\theta}_j} L) \tag{6}$$

If Δ exceeds a given threshold T, H_0 is rejected in favor of H_1. It is easy to show that:

$$\Delta = \min_{\boldsymbol{\theta}}[\frac{S_i(\boldsymbol{\theta})}{\sigma_i^2} + \frac{S_j(\boldsymbol{\theta})}{\sigma_j^2}] - \min_{\boldsymbol{\theta}} \frac{S_i(\boldsymbol{\theta})}{\sigma_i^2} - \min_{\boldsymbol{\theta}} \frac{S_j(\boldsymbol{\theta})}{\sigma_j^2} \tag{7}$$

Although in practice noise variance may show minor variation over the image, it is advantageous to assume it is constant over the whole image. That means $\sigma_i = \sigma_j = \sigma$. Then

$$\Delta = \frac{1}{\sigma^2}[\hat{S}_{ij} - \hat{S}_i - \hat{S}_j] \tag{8}$$

where

$$\hat{S}_i = \min_{\boldsymbol{\theta}} S_i(\boldsymbol{\theta}) \quad \text{and} \quad \hat{S}_{ij} = \min_{\boldsymbol{\theta}}[S_i(\boldsymbol{\theta}) + S_j(\boldsymbol{\theta})]$$

Hence, Δ is proportional to the increase in the residual sum of squares when a common model is chosen for the two regions instead of the optimal model for each one. This statistic yields our measure for similarity between motions of two regions R_i and R_j.

As an aside if we test the hypothesis: $H_0 : \boldsymbol{\theta}_i = \boldsymbol{\theta}_j$ AND $\sigma_i = \sigma_j$ versus $H_1 : \boldsymbol{\theta}_i \neq \boldsymbol{\theta}_j$ OR $\sigma_i \neq \sigma_j$ we then get the statistic used in [9]: $\Lambda = n_{ij}\log(\hat{S}_{ij}/n_{ij}) - n_i\log(\hat{S}_i/n_i) - n_j\log(\hat{S}_j/n_j)$. The drawback of Λ is that it is non-zero even if the motion parameters estimated in the two regions are equal.

For simplicity we now focus on the case where **f** is linear with respect to $\boldsymbol{\theta}$. If not, all results below are applicable if Taylor's approximations of **f** are made. We have shown that [6]:

$$\begin{aligned}\Delta(i,j) &= \frac{1}{2\sigma^2}(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_j)^\top \mathbf{H}_i(\mathbf{H}_i + \mathbf{H}_j)^{-1}\mathbf{H}_j(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_j) \\ &= \frac{1}{2\sigma^2}(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_j)^\top [\mathbf{H}_i^{-1} + \mathbf{H}_j^{-1}]^{-1}(\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_j)\end{aligned} \tag{9}$$

where $\mathbf{H}_i$ is the hessian matrix of $S_i(\boldsymbol{\theta})$, which is constant if **f** is linear. This matrix is a by-product of minimization of $S_i(\boldsymbol{\theta})$. It is important to note that $2\sigma^2\mathbf{H}_i^{-1}$ is the covariance matrix of $\hat{\boldsymbol{\theta}}_i$. Moreover, when a merging decision is accepted, the optimal motion parameters and the hessian of the union $R_i \cup R_j$ can be obtained directly from those of R_i and R_j as follows [6]:

$$\hat{\boldsymbol{\theta}}_{ij} = (\mathbf{H}_i + \mathbf{H}_j)^{-1}(\mathbf{H}_i\hat{\boldsymbol{\theta}}_i + \mathbf{H}_j\hat{\boldsymbol{\theta}}_j) \qquad \text{and} \qquad \mathbf{H}_{ij} = \mathbf{H}_i + \mathbf{H}_j \tag{10}$$

We also want to assess how close the motion of R_i is to a hypothesized motion model with parameters $\boldsymbol{\vartheta}$. In this case $\boldsymbol{\vartheta}$ is deterministic and its covariances are zero, eq. (9) then becomes:

$$\Delta_1(\hat{\boldsymbol{\theta}}_i, \boldsymbol{\vartheta}) = \frac{1}{2\sigma^2}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\vartheta})^\top \mathbf{H}_i(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\vartheta}) \tag{11}$$

This expression coincides with the Mahalanobis distance between $\hat{\boldsymbol{\theta}}_i$ and the class center $\boldsymbol{\vartheta}$.

B. Properties of the proposed measure

It is important to note that $\frac{1}{2\sigma^2}[(\mathbf{H}_i^{-1} + \mathbf{H}_j^{-1})]^{-1}$ in (9), the inverse of the covariance matrix of the vector $\hat{\boldsymbol{\theta}}_i - \hat{\boldsymbol{\theta}}_j$, is the *optimal scaling matrix*. This contrasts the ad-hoc one used in [8].

The other desirable properties of Δ are given in the following theorem, proofs of which can be found in [6]:

Theorem 1: The similarity measure $\Delta(i,j)$ defined in (9) satisfies the following properties:
1. $\Delta(i,j) > 0$; $\Delta(i,i) = 0$; $\Delta(i,j) = \Delta(j,i)$
2. Δ has a chi-squared distribution $\chi_p^2(\lambda)$ with degree of freedom p equal to the number of the motion parameters, whose non-centrality parameter is:

$$\lambda = \frac{1}{4\sigma^2}(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j)^\top \mathbf{H}_i(\mathbf{H}_i + \mathbf{H}_j)^{-1}\mathbf{H}_j(\boldsymbol{\theta}_i - \boldsymbol{\theta}_j) \tag{12}$$

3. For the velocity model (1) Δ is invariant to affine transformations of the coordinates in the image space.

The second property guarantees that Δ well discriminates regions undergoing the same motion from those undergoing different motions. In the former case $\boldsymbol{\theta}_i = \boldsymbol{\theta}_j$ and, hence, $\lambda = 0$ and Δ has a central chi-squared distribution. In the latter case $\lambda > 0$, the distribution of Δ becomes non-central and the measure tends to be large as the chance it exceeds the threshold T is much higher. The third property guarantees the independence of the merging result from translation or rotation of the coordinate system which is not the case for some earlier methods [8].

The proposed measure is not a metric as required by many standard clustering methods. It satisfies the conditions of positivity and symmetry as shown above, but not the triangular inequality. Thus, Δ is a good measure when an appropriate merging method is developed.

C. Merging procedure

Maximum likelihood approach for merging regions requires minimization of the residual sum of squares over all regions:

$$\mathbf{S}(\mathbf{z}, \boldsymbol{\vartheta}_1, \ldots, \boldsymbol{\vartheta}_K) = \sum_{i=1}^{M} S_i(\boldsymbol{\vartheta}_{z_i}) = \sum_{i=1}^{M} \sum_{\mathbf{x} \in R_i} |\mathbf{y}(\mathbf{x}) - \mathbf{f}(\mathbf{x}; \boldsymbol{\vartheta}_{z_i})|^2 \tag{13}$$

where $\mathbf{z} = (z_1, .., z_M)$ and $z_i \in \{1, \ldots, K\}$ is the region label specifying the index of the global moving surface to which R_i belongs. A similar approach was used in [2].

We now show that the above minimization can be viewed as a generalized version of standard K-means clustering and the algorithm used in [2] can be improved to save computations. Considering $\mathbf{f}$ is linear with respect to the motion parameters and applying Taylor's theorem, it is easy to show that:

$$S_i(\boldsymbol{\vartheta}_{z_i}) = \hat{S}_i + \frac{1}{2}(\hat{\boldsymbol{\theta}}_i - \boldsymbol{\vartheta}_{z_i})^\top \mathbf{H}_i (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\vartheta}_{z_i}) \tag{14}$$

Since $\hat{S}_i$ is constant, the minimization of $\mathbf{S}$ boils down to minimization of $\mathbf{S}'$ where:

$$\begin{aligned} \mathbf{S}'(\mathbf{z}, \boldsymbol{\vartheta}_1, .., \boldsymbol{\vartheta}_K) &= \frac{1}{2} \sum_{i=1}^{M} (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\vartheta}_{z_i})^\top \mathbf{H}_i (\hat{\boldsymbol{\theta}}_i - \boldsymbol{\vartheta}_{z_i}) \\ &= \sigma^2 \sum_{i=1}^{M} \Delta_1(\hat{\boldsymbol{\theta}}_i, \boldsymbol{\vartheta}_{z_i}) \end{aligned} \tag{15}$$

This is, in fact, similar to the objective function in the K-means algorithm except that the Euclidean distance between $\hat{\boldsymbol{\theta}}_i$ and $\boldsymbol{\vartheta}_{z_i}$ is replaced with the Mahalanobis distance (11). Like traditional K-means, good initial cluster centers are required. The ad-hoc pixel-based technique used in [2] for deriving the initial set of global motions is very likely to miss the motion of foreground objects. We propose an algorithm encompassing two stages where the first one performs hierarchical region merging and provides a good starting point for the generalized K-means iterative procedure in the second:

MERGING STAGE 1

1. Specify K, the number of objects. Initialize each cluster with one region: $C_m = \{R_m\}; m = 1, ..., M$.
2. Merge the two adjacent clusters C_i and $C_j$with the smallest Δ defined in (9). Repeat until the number of clusters is reduced to K.

MERGING STAGE 2

1. Assign each $\hat{\boldsymbol{\theta}}_i$ to the nearest cluster center $\boldsymbol{\vartheta}_{z_i}$, using the distance measure Δ_1, defined in (11).
2. Update the cluster centers $\boldsymbol{\vartheta}_k$ so that the sum of differences between the center and the cluster members is minimized. As we showed in [6] the new $\boldsymbol{\vartheta}_k$ can be found by solving the following system of equations:

$$(\sum_{z_i=k} \mathbf{H}_i)\boldsymbol{\vartheta}_k = \sum_{z_i=k} \mathbf{H}_i \hat{\boldsymbol{\theta}}_i \tag{16}$$

3. Repeat 1 and 2 until cluster membership is unchanged.

Since the number of objects is specified in the beginning, the value of variance of noise does not affect the final result and we do not have to specify σ. The convergence in the second stage is guaranteed as the cost function always decreases. Actually, if starting from the same initial set of cluster centers, the second stage gives the same result as the two-step iterative procedure used in [2] does [1]. However, our algorithm requires much less computations, considering that $\hat{\boldsymbol{\theta}}_i$ and $\mathbf{H}_i$ are already obtained from the least-squares minimization of S_i. Note also, that the above algorithm merges non-adjacent regions as well, which is not the case for some methods [1], [9], [5].

IV. Results

In Figures 2 and 3 we show the results of applying the proposed merging algorithm on standard test sequences. The initial segmentation was obtained with the morphological multiscale technique [7]. The results for *Table Tennis* and *Flower Garden* were obtained with optic flow matching. As measurements $\mathbf{y}(\mathbf{x})$ we used the dense optic flow field, computed from two successive images using the hierarchical method in [3]. For the *Calendar* sequence we used the model, which is based on the linearized intensity matching equation [6].

The improvements due to the use of the new similarity measure are confirmed by comparison with Figure 2c,d, in which we show the results of the existing methods in [9] and [2] applied for the same set of initial regions. More elaborate evaluation can be found in [6].

V. Conclusion

We have proposed a new criterion for similarity of regions movement in a video scene based on a statistical test for equality of motion parameters. The uncertainty in parameter estimation is incorporated in an optimal way. Using this measure we have developed a new merging algorithm consisting of two stages. The agglomerative merging in the first stage provides a good starting point for the second stage in which the regions are merged according to a K-means like algorithm. The improved performance over existing methods has been demonstrated on real sequences. As extracted objects and their motion parameters are accurate, they can be used for content-based video retrieval in digital libraries.

References

[1] G. Adiv, "Determining 3d motion and structure from optical fbws generated by several moving objects," *IEEE Trans. on PAMI*, vol. 7, no. 4, pp. 384–401, 1985.

[1] We note, by the way, that in [2, section 4.3] the mixed use of intensity matching and optic fbw matching dose not guarantee convergence.

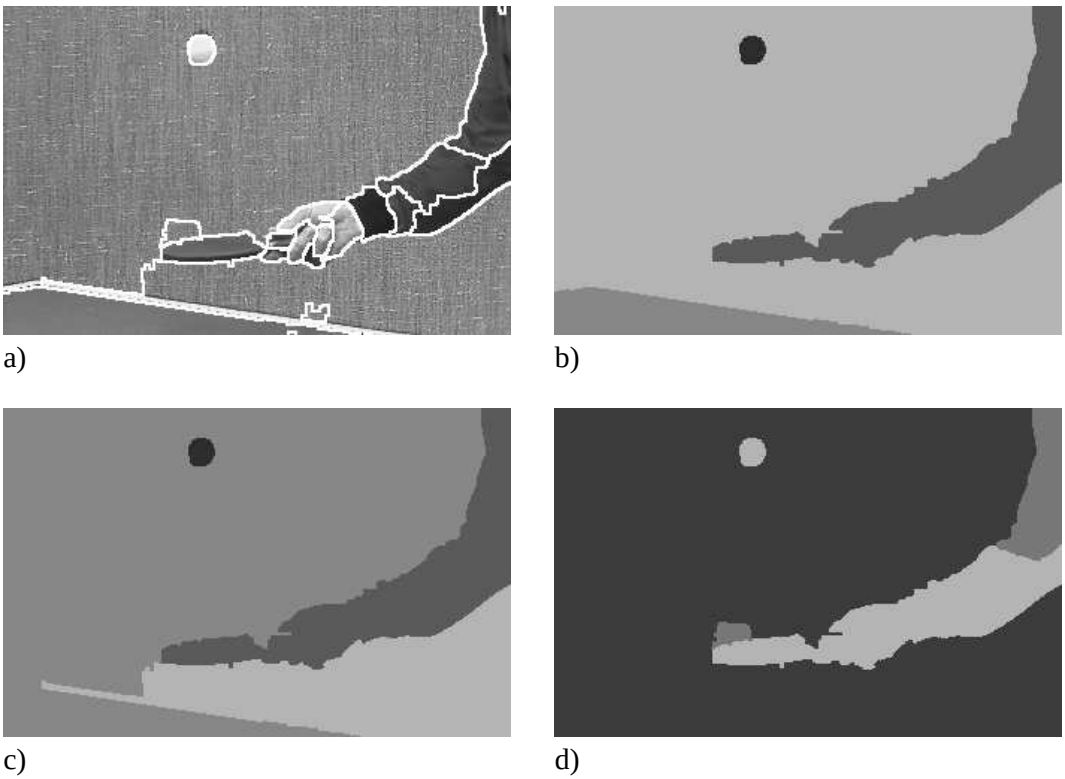

Fig. 2. *a) One frame from* Table Tennis *and the initial segmentation, consisting of 23 regions; b) Merging result of the proposed method, K=4; c) Result of the method in [9]; d) Result of the method in [2].*

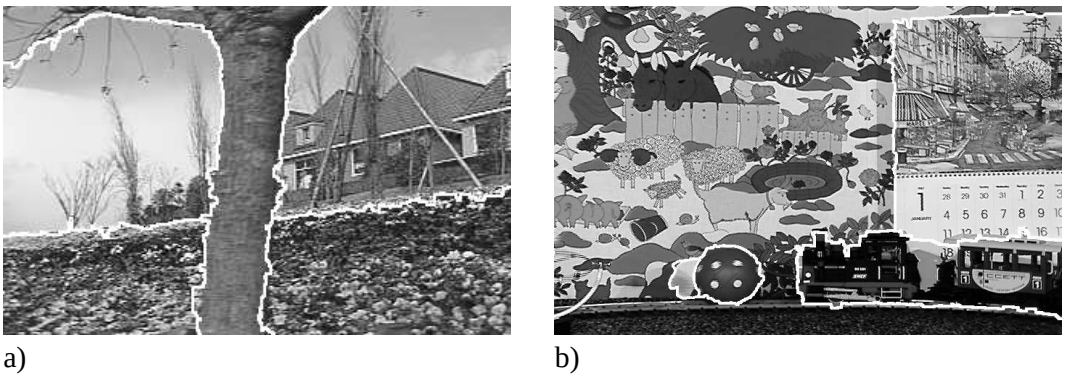

Fig. 3. *Merging result of the proposed method for a)* Flower Garden, *K=3; and b)* Calendar, *K=5. Object boundaries are marked by white lines.*

[2] Y. Altunbasak, P.E. Eren, and A.M. Tekalp, "Region-based parametric motion segmentation using color information," *Graphical Models and Image Proc.*, vol. 60, no. 1, pp. 13–23, Jan. 1998.

[3] J.R. Bergen, P. Anandan, K.J. Hanna, and R. Hingorani, "Hierarchical model-based motion estimation," in *Proc. 2nd European Conf. of Comp. Vision*, 1992, pp. 237–252.

[4] P.J. Huber, *Robust statistic*, John Wiley, New York, 1981.

[5] F. Moscheni, S. Bhattacharjee, and M. Kunt, "Spatiotemporal segmentation based on region merging," *IEEE Trans. on PAMI*, vol. 20, no. 9, pp. 897–915, Sept. 1998.

[6] H.T. Nguyen, M. Worring, and A. Dev, "Robust motion-based segmentation in video sequences," Tech. Rep. 4, Intell. Sensory Inf. Sys. Group, Univ. of Amsterdam, Dec. 1998, available at http://carol.wins.uva.nl/˜rein/isis.html.

[7] P.Salembier, "Morphological multiscale segmentation for image coding," *Signal Processing*, vol. 38, no. 3, pp. 359–386, Sept. 1994.

[8] J.Y.A. Wang and E.H. Adelson, "Representing moving images with layers," *IEEE Trans. on Image Proc.*, vol. 3, no. 5, pp. 625–628, Sept. 1994.

[9] L. Wu, J. Benois-Pineau, Ph. Delagnes, and D. Barba, "Spatio - temporal segmentation of image sequences for object-oriented low bit-rate image coding," *Signal Proc. : Image Comm.*, vol. 8, pp. 513–543, Sept. 1996.