Pattern Recognition, 32(10):1751-1766 (1999)

Segmentation from Motion: Combining Gabor- and Mallat-Wavelets to Overcome the Aperture and Correspondence Problems

Laurenz Wiskott*
Institut für Neuroinformatik
Ruhr-Universität Bochum
D–44780 Bochum, Germany
http://www.neuroinformatik.ruhr-uni-bochum.de

Abstract

A new method for segmentation from motion is presented, which is designed to be part of a general object-recognition system. The key idea is to integrate information from Gabor- and Mallat-wavelet transforms of an image sequence to overcome the aperture and the correspondence problem. It is assumed that objects move fronto-parallel. Gabor-wavelet responses allow accurate estimation of image flow vectors with low spatial resolution. A histogram over this image flow field is evaluated and its local maxima provide a set of *motion hypotheses*. These serve to reduce the correspondence problem occurring in utilizing the Mallat-wavelet transform, which provides the required high spatial resolution in segmentation. Segmentation reliability is improved by integration over time. The system can segment several small, disconnected, and openworked objects, such as dot patterns. Several examples demonstrate the performance of the system and show that the algorithm behaves reasonably well, even if the assumption of fronto-parallel motion is not met.

Keywords: segmentation from motion, Gabor-wavelet transform, Mallat-wavelet transform, integration, motion hypotheses.

1 Introduction

There is a large literature on systems for segmentation from motion [1, 2]. Most of the systems follow the approach of Fennema and Thompson [3] and Horn and Schunck [4] and are based on the spatio-temporal image gradient [5, 6, 7]. For image location $\mathbf{x} = (x, y)$ and time t in an image sequence $I(x, y, t)$, the motion constraint equation is $\frac{\partial I}{\partial x}\frac{dx}{dt} + \frac{\partial I}{\partial y}\frac{dy}{dt} + \frac{\partial I}{\partial t} = 0$, where $\dot{\mathbf{x}}$ is the local image flow vector. This equation yields a local estimate of the image flow, but only in the direction of the gradient. Several estimates must be somehow combined to estimate the true local image flow, usually oblique to the gradient. Due to noise and the inherent aperture problem, i.e. the fact that only the component of the image flow vector in the direction of the gradient can be estimated, the estimated image flow field is unreliable. In addition, this method is restricted to slow motion or small displacements, because of the local linear approximation of the grey value distribution implicit in the motion constraint equation. Segmentation methods based on this image flow estimation often partition an image into different segments and fit a set of motion models to the flow vectors of each segment at the same time, optimizing the partitioning and the motion model parameters simultaneously [8, 9]. These systems provide a full field segmentation, possibly suppressing regions with insufficient evidence for motion. An interesting variant of these systems does not segment the image into distinct regions, but separates two overlaying components, as may occur for transparent objects [5].

Another common class of systems for segmentation from motion is based on matching feature points, such as corners or interest points [10, 11, 12]. Since these systems process only a relatively sparse set of

[*] Corresponding author: Institute for Advanced Studies, Wallotstr. 19, D-14193 Berlin, Germany, wiskott@wiko-berlin.de.

feature points, they are used to detect and track moving objects in a scene, rather than segmenting them with high resolution. Such a method is also useful if 3D-structure is to be revealed from motion, because fairly exact point-to-point correspondences are established over a sequence of frames. Instead of matching feature points, some systems match small image blocks [13, 14]. These systems are preferably used in the context of low-bit-rate video coding. This method again results in a rather crude segmentation with a resolution given by the block-size. However, the purpose of video coding is in any case compression, rather than segmentation. An interesting variant of this matching method first performs a segmentation on each single frame, based on grey value differences, and then matches the resulting regions [15]. This method provides full field segmentation with high resolution. These matching-based methods do not suffer from the aperture problem. However, they have the correspondence problem instead, which means that the features may not be discriminative enough to disambiguate different matching alternatives.

The segmentation method presented here differs from these two classes and has been motivated in the context of a larger effort to build a general system for object recognition [16, 17]. This object-recognition system uses labeled graphs to represent objects and elastic graph matching to compare graphs with images. The nodes of the graphs are labeled with Gabor-wavelet responses, though other types of filter responses are possible as well. In developing this system, we follow some biologically motivated design principles. One of them is to use general purpose representations rather than representations specific for a particular object class. Others are the principles of robustness and flexibility. The work presented here is an attempt to develop a segmentation method which follows similar concepts and joins smoothly with this general object recognition system.

Another important principle of biological systems is that of integration. Natural systems have often to deal with incomplete, ambiguous, and noisy data and tend to integrate several cues to come to a reliable percept. Algorithmic systems, on the other hand, tend to focus on one particular method and to optimize it as far as possible, neglecting other cues or methods which could supplement the existing approach. Integration is particularly important for segmentation, and algorithmic segmentation systems actually take advantage of integration at two levels. Firstly, since local image flow estimation and matching techniques are noisy and unreliable due to the aperture and the correspondence problems, respectively, the systems have to integrate information to achieve a reasonable level of performance. This can be done by simple averaging or by other more sophisticated methods, such as robust estimation with outlier detection [18, 19]. Integration has been applied in the spatial domain [4], scale space [20], and time [6]. Secondly, there exist some examples in which different segmentation cues, such as color and motion, are integrated [21, 22] or where segmentation cues are combined with object recognition [23], but these systems are the exception rather than the rule. These two types of integration are of low and high level, respectively.

This work investigates a third, intermediate level of integration where different techniques for the same cue are combined. In a first stage, Gabor-wavelets are used for image flow estimation. They have large supports and are sensitive to texture components of different orientations, such that they do not have the aperture problem. But they provide only a coarse image flow field with low spatial resolution. From this image flow field, the system extracts motion hypotheses, i.e. a small set of dominant image flow vectors. This small set of motion hypotheses is used in a second stage to constrain possible matches for edges detected by the Mallat-wavelets to few small subregions, thus reducing the correspondence problem significantly. In addition, temporal integration helps to further disambiguate matches. The main conceptual aspects of this work can be summarized as follows:

General purpose preprocessing: The preprocessing stages used here, Gabor- and Mallat-wavelet transforms, are also useful for object representation [24, 25]. Expensive, task-specific representations which are useless for other purposes are avoided. This is important, because the final goal should be to integrate segmentation and object recognition into one system.

Biologically plausible preprocessing: Beside their technical advantages, Gabor- and Mallat-wavelet transforms are also biologically plausible, as shown by physiological [26, 27] as well as psychophysical experiments [28].

Simple motion model: While most systems use affine, 3D-rigid, or similar motion models, the system presented here uses the simplest motion model, the fronto-parallel translation. Psychophysical experiments provide evidence that this is the dominant motion model in human subjects [29] and that humans perceive moving patterns not necessarily according to a 3D-rigid motion model [30].

No restrictive object model: In contrast to many other approaches, no spatial regularization is applied. No assumption is made about the shape of the objects, e.g. compactness [31, 12], or their number, e.g. one object in front of a resting background [32, 33]. This is to keep the system as general as possible. However, the method presented generates an intermediate representation which is suitable to incorporate regularization constraints easily.

Segmentation on edges only: As a consequence of the previous aspect, segmentation cannot be enforced for each pixel but only for those which provide enough evidence. Therefore, the segmentation task is here restricted to edge pixels, carrying more information than others. This is in part motivated by work done by Mallat and Zhong [25] which has shown that images can be reconstructed from edge information.

Integrate information from different kernels: The key idea of this work is to reduce the aperture and correspondence problems by combining the two different preprocessing representations. Based on the Gabor-wavelet transform, image flow vectors can be accurately computed at low spatial resolution. This stage is used to form *motion hypotheses*. These are then used to reduce the correspondence problem for the Mallat-wavelet responses, which provide high spatial resolution.

Integration over time: While spatial regularization would explicitly or implicitly introduce a restricting object model, temporal regularization can be employed without limiting the applicability of the method. Any object to be segmented has to persist in time over several frames. Therefore, segmentation evidence can be integrated over time. This does not require any continuity in the motion itself. Objects may jump back and forth from frame to frame.

The system developed is explained in detail in the following section. Section 3 shows further examples of segmented image sequences and will demonstrate the performance of the system. Section 4 gives a broader picture and a comparison with other systems.

2 The System

2.1 Segmentation on Two Frames

The method for two frames, i.e. without temporal integration, can be summarized in four stages:

1. Based on the Gabor wavelet transform, the image flow field between the two images is computed. Since the Gabor wavelets have a large support, this method does not have a serious aperture or correspondence problem, but the spatial resolution of the computed flow field is low.

2. It is assumed that all objects in the scene move mainly translationally, i.e. fronto-parallel. A histogram over the flow vectors is evaluated and its peaks provide accurate information on the motion vectors of the different objects, called *motion hypotheses*. The spatial information of the image flow field is disregarded.

3. For each edge pixel, each motion hypothesis is checked. The local grey-value gradient of an edge in one frame is compared with the gradient of an edge in the next frame, but taken from a pixel which is displaced according to the motion hypothesis under consideration. This comparison provides an *accordance* value for each hypothesis, which is high if the gradients are similar and low if they are not.

4. Finally, each edge pixel is categorized in the motion hypothesis class for which it has achieved the highest accordance value. This is done for each edge pixel individually.

In the following, the four stages are explained in greater detail. In Section 2.2, the method will be extended to a sequence of frames, for which the segmentation can be improved by integration over time.

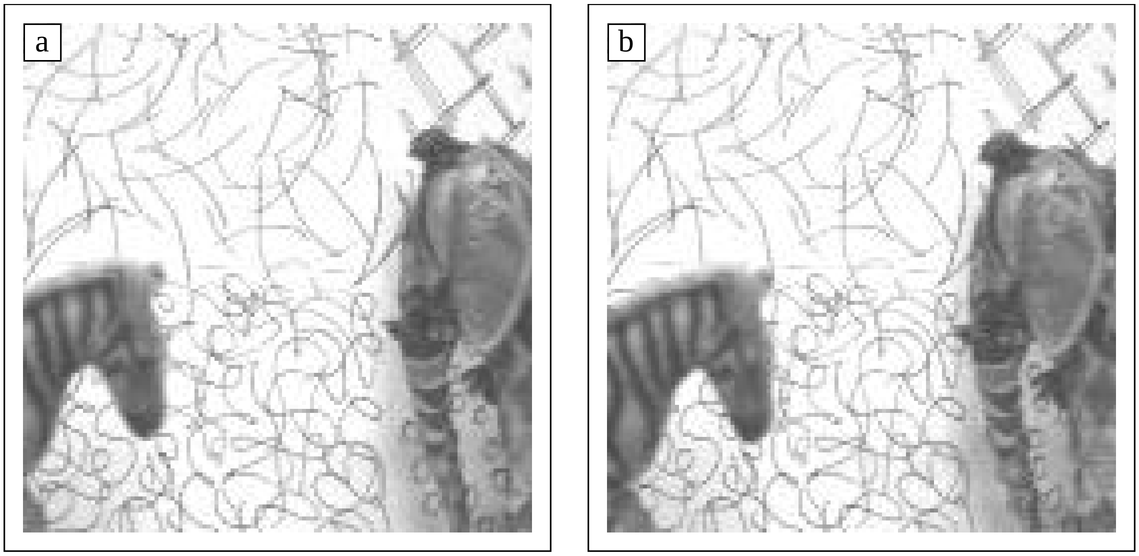

Figure 1: **(a)** Frame 14 and **(b)** Frame 15 of the moving-animals sequence. The zebra moves fast to the right, the elephant slowly to the left. The images have a resolution of 128 × 128 pixels with 256 grey values.

2.1.1 Image Flow Estimation

The first stage of the method is the computation of an image flow field. The method used here is based on a convolution with *Gabor wavelets*, which have the shape of localized, plane waves. This Gabor wavelet transform yields complex coefficients with phases usually varying with the main frequency of the corresponding kernels. The distance $\mathbf{d}$ between two points and the phase difference $\Delta\phi$ of the coefficients at these two points have a relationship approximately like $\Delta\phi = \mathbf{d} \cdot \mathbf{k}$, where $\mathbf{k}$ is the wave vector of the kernel's main frequency and $\cdot$ indicates the usual dot product of two vectors. Using several kernels of different orientations can provide an accurate estimate of $\mathbf{d}$ based on the several $\Delta\phi$. Similarly, phase differences between pairs of coefficients at the same location in two successive frames can be used to estimate the translation of the underlying gray-value distribution at that location. Doing this at each pixel provides an image flow field.

Gabor Wavelet Transform The Gabor wavelet transform [17, 24] for an image $\mathcal{I}(\mathbf{x})$ is defined as a convolution

$$\mathcal{J}_j(\mathbf{x}) = \int \mathcal{I}(\mathbf{x}')\psi_j(\mathbf{x}-\mathbf{x}')d\mathbf{x}' \tag{1}$$

with a family of Gabor wavelets

$$\psi_j(\mathbf{x}) = \frac{k_j^2}{\sigma^2}\exp\left(-\frac{k_j^2 x^2}{2\sigma^2}\right)\exp(i\,\mathbf{k}_j\cdot\mathbf{x}) \tag{2}$$

in the shape of plane waves with wave vector $\mathbf{k}_j$, restricted by a Gaussian envelope function. k_j and x indicate the norm of vectors $\mathbf{k}_j$ and $\mathbf{x}$. Analogous definitions hold for other vectors throughout the paper. We employ a discrete set of 5 different frequencies, index $\nu = 0, ..., 4$, and 8 orientations, index $\mu = 0, ..., 7$,

$$\mathbf{k}_j = \begin{pmatrix} k_{jx} \\ k_{jy} \end{pmatrix} = \begin{pmatrix} \kappa_\nu \cos\varphi_\mu \\ \kappa_\nu \sin\varphi_\mu \end{pmatrix},\ \kappa_\nu = 2^{-\frac{\nu+2}{2}}\pi,\ \varphi_\mu = \mu\frac{\pi}{8}, \tag{3}$$

with index $j = \mu + 8\nu$. k_{jx} and k_{jy} indicate the x- and y-component of wave vector $\mathbf{k}_j$. This sampling evenly covers a band in frequency space. The width σ/k_j of the Gaussian is controlled by the parameter $\sigma = 2\pi$. One speaks of a wavelet transform, since the family of kernels is self-similar, all kernels being generated from one *mother wavelet* by dilation and rotation.

A *jet* $\mathcal{J}$ is defined as the set $\{\mathcal{J}_j\}$ of 40 complex coefficients obtained for one image point. It can be written as

$$\mathcal{J}_j = a_j \exp(i\phi_j) \tag{4}$$

with amplitudes $a_j(\mathbf{x})$, which slowly vary with position, and phases $\phi_j(\mathbf{x})$, which rotate with a rate approximately set by the spatial frequency of wave vector $\mathbf{k}_j$.

Gabor wavelets are technically interesting for image representation [24] and object recognition [17]. They are also biologically plausible, because they have a shape similar to the receptive fields of simple cells found in the visual cortex of vertebrate animals [26, 27].

Estimating Flow Vectors Estimating image flow vectors for successive frames of a sequence, disparities between pairs of stereo images, or small positional displacements in a matching procedure are closely related tasks. The method presented here was adopted from a stereo algorithm [34] and has also been used for matching purposes [17].

Consider two jets $\mathcal{J}$, $\mathcal{J}'$ taken from the same pixel position in two successive frames. The underlying object in the scene may have moved by a vector $\mathbf{d}$. The phases of the jet coefficients $\mathcal{J}_j$ then vary mainly according to their corresponding wave-vectors $\mathbf{k}_j$, yielding phase differences $\Delta\phi_j = \phi_j - \phi'_j \approx \mathbf{d}\cdot\mathbf{k}_j$. Vice versa, since the phases of $\mathcal{J}$ and $\mathcal{J}'$ are known, the displacement $\mathbf{d}$ can be estimated by matching the terms $\mathbf{d}\cdot\mathbf{k}_j$ with the phase differences $\Delta\phi_j$, which can be done by maximizing the function

$$\mathcal{S}_\phi(\mathcal{J},\mathcal{J}') = \frac{\sum_j a_j a'_j \cos(\Delta\phi_j - \mathbf{d}\cdot\mathbf{k}_j)}{\sqrt{\sum_j a_j^2 \sum_j a'^2_j}} \tag{5}$$

with respect to $\mathbf{d}$. Sums over j run over all 40 coefficients, i.e. $j = 0, ..., 39$. The phase difference terms $\cos(\Delta\phi_j - \mathbf{d}\cdot\mathbf{k}_j)$ are weighted by the amplitudes a_j and a'_j, because strongly responding jet coefficients are expected to provide more reliable phases. The denominator normalizes $\mathcal{S}_\phi$ to the range of $[-1, ..., 1]$, so that it is independent of the contrast of the images. $\mathcal{S}_\phi$ can be used as a similarity function [17] and provides an indication of how reliable the displacement estimation is.

The function $\mathcal{S}_\phi$ is maximized in its Taylor expansion

$$\mathcal{S}_\phi(\mathcal{J},\mathcal{J}') \approx \frac{\sum_j a_j a'_j [1 - 0.5\,(\Delta\phi_j - \mathbf{d}\cdot\mathbf{k}_j)^2]}{\sqrt{\sum_j a_j^2 \sum_j a'^2_j}} \,. \tag{6}$$

Setting $\frac{\partial}{\partial d_x}\mathcal{S}_\phi = \frac{\partial}{\partial d_y}\mathcal{S}_\phi = 0$ and solving for $\mathbf{d}$ leads to

$$\mathbf{d}(\mathcal{J},\mathcal{J}') = \begin{pmatrix} d_x \\ d_y \end{pmatrix} = \frac{1}{\Gamma_{xx}\Gamma_{yy} - \Gamma_{xy}\Gamma_{yx}} \times \begin{pmatrix} \Gamma_{yy} & -\Gamma_{yx} \\ -\Gamma_{xy} & \Gamma_{xx} \end{pmatrix} \begin{pmatrix} \Phi_x \\ \Phi_y \end{pmatrix}, \tag{7}$$

if $\Gamma_{xx}\Gamma_{yy} - \Gamma_{xy}\Gamma_{yx} \neq 0$, with

$$\begin{aligned} \Phi_x &= \sum_j a_j a'_j k_{jx} \Delta\phi_j, \\ \Gamma_{xy} &= \sum_j a_j a'_j k_{jx} k_{jy}, \end{aligned}$$

and $\Phi_y, \Gamma_{xx}, \Gamma_{yx}$, and Γ_{yy} are defined correspondingly.

Eq. (7) yields a straightforward method for estimating the displacement vectors for two jets taken from the same pixel positions in two successive frames. Without further modifications, i.e. with $j = 0, ..., 39$, it can determine displacements up to half the wavelength of the highest frequency kernel, which would be two pixels for $\kappa_0 = \pi/2$. The range can be increased by using low frequency kernels only, $j = 32, ..., 39$. For the largest kernels, the estimated displacement can be 8 pixels. This estimate can be used to correct the phases of the higher frequency coefficients by multiples of 2π, since phases are cyclic variables and determined only up to an arbitrary integer multiple of 2π. One can then repeat the procedure by taking into account also the next higher frequency level, $j = 24, ..., 39$, and refining the result. Repeating this procedure until all levels are used, $j = 0, ..., 39$, is the schedule which is used here. Thus all five levels are used, but estimation of displacements up to 8 pixels is possible. (If the displacement is large, it may be reasonable to discard some of the highest levels, but, for simplicity ,this is not done here.) The estimated image flow field for the moving-animals sequence in Fig. 1 is shown in Fig. 2.

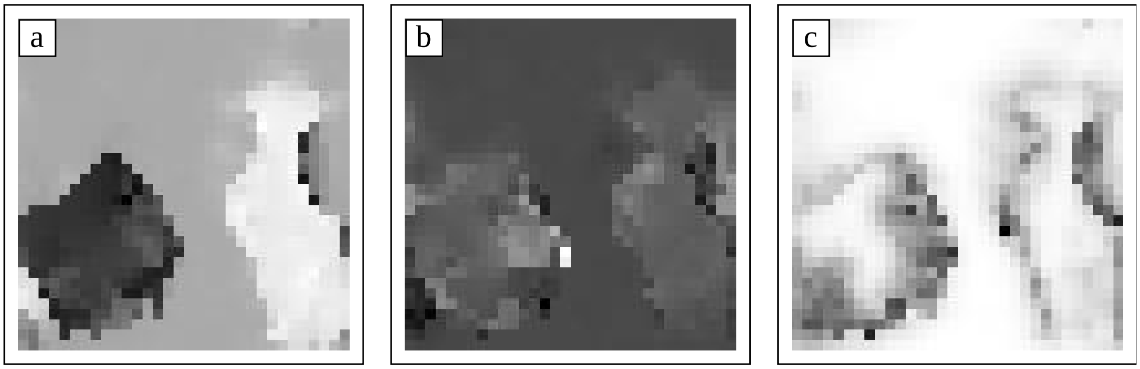

Figure 2: **(a)** x-component and **(b)** y-component of the image flow field for Frames 14 and 15 shown in Fig. 1. Negative values are light and positive values dark. The x-axis goes horizontally from left to right, the y-axis vertically from bottom to top. The flow field was computed at a lower resolution of 32 × 32 blocks of 4 × 4 pixels. It can be seen that the flow vectors are fairly constant within regions, indicating high accuracy of the estimated displacements. The borders of the regions, on the other hand, coincide only roughly with the borders of the objects. Thus the spatial resolution of this image flow algorithm is low. **(c)** An error measure $(1 - \mathcal{S}_\phi)$ indicates where the flow vectors are unreliable (dark regions). Notice that the Gabor-transform, and therefore also the image flow field, was computed with wrap-around conditions.

The key point here is that, due to the large support of the Gabor wavelets, this image flow estimation does not have an aperture or correspondence problem, except in some degenerated cases. The estimated image flow vectors are thus usually accurate with respect to the estimated displacements. However, the spatial resolution is low for the same reason. The spatial information is therefore disregarded and only a histogram over the image flow field is evaluated in order to obtain accurate hypotheses about the different displacements dominating in the scene. This also helps eliminating errors in image flow estimation by averaging. Ideally, each moving object leads to one hypothesis, plus one hypothesis for the background.

Image flow estimation may fail for three reasons. Firstly, the motion may exceed the limit of eight pixels per frame. This is a hard limit and the frame rate needs to be high enough to avoid this. Secondly, if image texture varies over a large region only in one dimension, e.g. in case of a pure sine grating, the system is faced with a macroscopic aperture problem and the image flow component perpendicular to the direction of variation cannot be estimated. This failure can be detected because $\Gamma_{xx}\Gamma_{yy} - \Gamma_{xy}\Gamma_{yx}$ would have a value close to zero. Since the macroscopic aperture problem is a rare event in real image sequences, no special procedure has been implemented in the system to deal with this situation. Thirdly, due to occlusions, new image information may appear or old information may disappear, in which case the Gabor wavelet representation may vary significantly even for corresponding locations. Therefore, image flow vectors with too low a similarity value $\mathcal{S}_\phi$ are disregarded. By looking at the values of $\Gamma_{xx}\Gamma_{yy} - \Gamma_{xy}\Gamma_{yx}$ and $\mathcal{S}_\phi$, one can also adapt the number of levels used locally to achieve higher resolution and more reliable image flow estimation. However, due to the averaging in the motion histogram, no such elaborated procedure seems necessary. These kinds of special cases have been treated more explicitly elsewhere [35].

2.1.2 Motion Hypotheses

The second stage of the method is concerned with the extraction of motion hypotheses from the image flow field. Since it is assumed that objects move mainly translationally, this can simply be done by detecting local maxima in the image flow histogram, a histogram over the flow field vectors; see Fig. 3. (Actually, the histogram is restricted to edge pixels, as will be explained later.) Examples violating the assumption of fronto-parallel motion are given in Section 3. In order to avoid detecting too many irrelevant maxima, a low-pass filter is applied to the histogram and only maxima of a certain minimal height relative to the histogram maximum are accepted. The result of this stage is usually a small number of displacement vectors $\mathbf{v}_n$ representing frequently occurring flow vectors, e.g. $\mathbf{v}_3$ corresponding to the background, $\mathbf{v}_0$ and $\mathbf{v}_2$ corresponding to two differently moving objects in the scene, and $\mathbf{v}_1$ being one spurious motion hypothesis. (The indices do not refer to any particular order.) A displacement $\mathbf{d}$ can be compared with a motion hypothesis $\mathbf{v}_n$ by the displacement similarity function

$$\mathcal{S}_d\left(\mathbf{v}_n, \mathbf{d}\right) = \max\left\{1 - \frac{\left(\mathbf{v}_n - \mathbf{d}\right)^2}{r^2}, 0\right\} \tag{8}$$

with a parameter r to set its sensitivity.

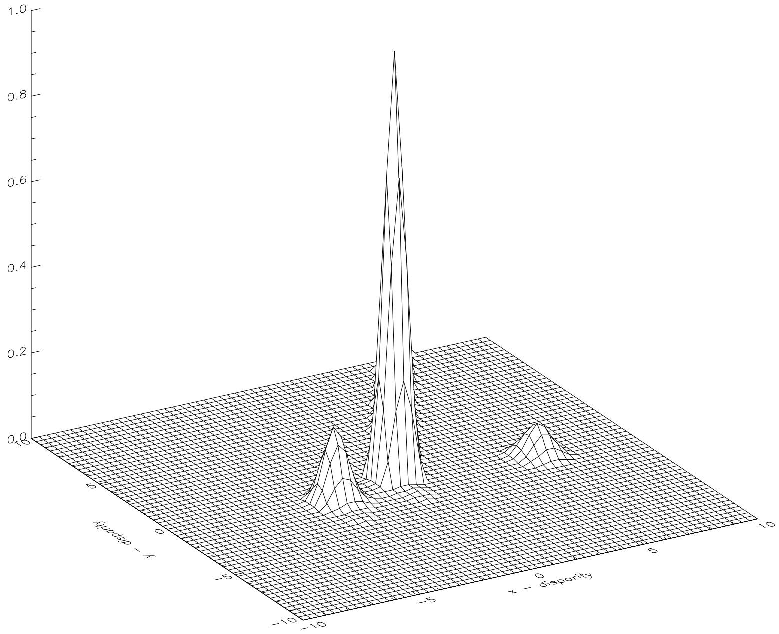

Figure 3: Image flow histogram for Frames 14 and 15 of the moving-animals sequence. It shows three prominent local maxima. The largest one corresponds to the resting background. The smaller ones on the left and right side correspond to the elephant and zebra, respectively. The zebra moves faster to the right than the elephant to the left. In addition there is a fourth local maximum, too small to be visible here and not corresponding to any object.

The advantage of the motion hypotheses is that they drastically reduce the correspondence problem for an image flow estimation algorithm based on more localized features, such as edges. Instead of testing all possible displacements within a certain distance range, only a few of them suggested by the motion hypotheses need to be taken into consideration; see Fig. 4.

2.1.3 Edge Valuation

The third stage of the method uses the Mallat-wavelet transform [25], which can be thought of as the grey-value gradient at different levels of resolution. This stage is similar to a matching algorithm and is therefore faced with the correspondence problem, which is particularly severe for such a simple feature as the local gradient. Two methods are employed here to bypass the correspondence problem. Firstly, the evaluation of the Mallat-wavelet transform is restricted to edges, as defined by the modulus maxima [25]. Edges have a particularly high information content and are less ambiguous than gradients in general. Secondly, the system does not match the edges between two frames, but evaluates their *accordance* with the different motion hypotheses obtained in stage two. This reduces the correspondence problem significantly, but leaves

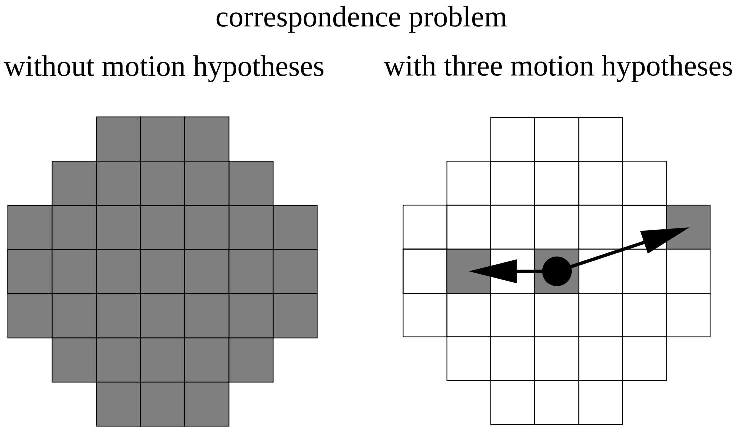

Figure 4: Illustration of the reduced correspondence problem: Without motion hypotheses, a whole region within a certain diameter has to be tested for possibly corresponding pixel locations with a similar grey value gradient. With motion hypotheses, the regions to consider are reduced to a few small spots.

it partially unresolved at the same time, because the final decision between the small number of motion hypotheses is not yet made.

It is important to notice that going from a Gabor-wavelet representation to a Mallat-wavelet representation is qualitatively different from going from one level in a resolution hierarchy to a higher resolution level. In a conventional resolution hierarchy, only the scale of the representation changes from one level to the next, but not the quality, thus the problems remain the same. For instance, an edge-based method would have the correspondence problem on any level. The Mallat-wavelets, on the other hand, are of different quality than the Gabor-wavelets. Furthermore, the algorithms change with the representation. Thus the strengths and weaknesses are different and can complement each other, even on the same level of frequency-resolution. Thus the combination of a Gabor-wavelet representation and a Mallat-wavelet representation is not comparable to a conventional resolution hierarchy.

In the following, the considerations and examples will be restricted to the highest level of the Mallat-wavelet transform, which provides edges with a single pixel resolution; see Fig. 5. The terms gradient and edges will be used instead of wavelet transform and modulus maxima, respectively, because they are more common and denote basically the same. However, as will be discussed later, it is important from a conceptional point of view that the modulus maxima of the wavelet transform provide enough information to reconstruct images [25].

Given the image $\mathcal{I}(\mathbf{x})$, the gradient is defined as

$$\mathbf{g}(\mathbf{x}) = (\partial \mathcal{I}/\partial x, \partial \mathcal{I}/\partial y) = (g\cos\theta, g\sin\theta), \tag{9}$$

with $g = g(\mathbf{x}) \geq 0$ denoting the magnitude and $\theta = \theta(\mathbf{x})$ the direction angle of $\mathbf{g}$. Two gradients are compared by the gradient similarity function

$$\mathcal{S}_g(\mathbf{g}, \mathbf{g}') = \begin{cases} \min\{\frac{g}{g'}, \frac{g'}{g}\}\cos^2(\theta - \theta') & \text{for } |\theta - \theta'| < \pi/2 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

Each edge pixel is tested for each motion hypothesis to determine whether its grey-value gradient is similar to a gradient of an edge pixel in the other frame at a position shifted by the flow vector of the considered motion hypothesis. To be more robust, a small neighborhood is tested and the *accordance function* is defined as the maximum which can be found by varying $\mathbf{d}$:

$$\mathcal{A}(\mathbf{x}, t, \mathbf{v}_n) = \max_{\mathbf{d}} \left\{ \mathcal{S}_d\left(\mathbf{v}_n, \mathbf{d}\right)\, \mathcal{S}_g\left(\mathbf{g}(\mathbf{x}, t), \mathbf{g}(\mathbf{x} - \mathbf{d}, t - 1)\right) \right\}, \tag{11}$$

where t indicates the current frame. Notice that only edge pixels are evaluated and taken into consideration. This accordance function defines an accordance map for each motion hypothesis, shown in Fig. 6, indicating whether an edge might have moved by the respective flow vector $\mathbf{v}_n$ or not. It is important to note that the accordance maps may still contain ambiguities as to which edge fits which motion hypothesis. This would be expressed by high accordance values in more than one accordance map for a single edge pixel.

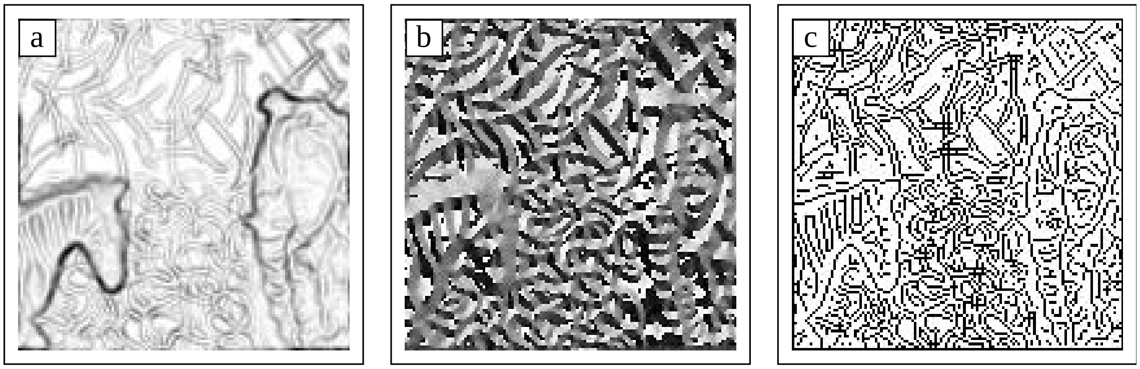

Figure 5: Local gradient as an edge representation. **(a)** the amplitude of the gradient from zero (white) to maximum (black), **(b)** its direction angle between 0 (black) and 2π (white), and **(c)** the edges as defined by the modulus maxima (black).

Edges at the outlines of objects are often subject to particularly large changes, because the background may change quickly. This problem potentially occurs at all occlusion boundaries. However, since the orientation of edges does not change, the similarity between corresponding edges may still be greater than zero as long as the gradient does not reverse. Furthermore, it is not the absolute similarity value of edges that determines the segmentation but relative values, so that the system is robust to some extent to changes at moving occlusion boundaries.

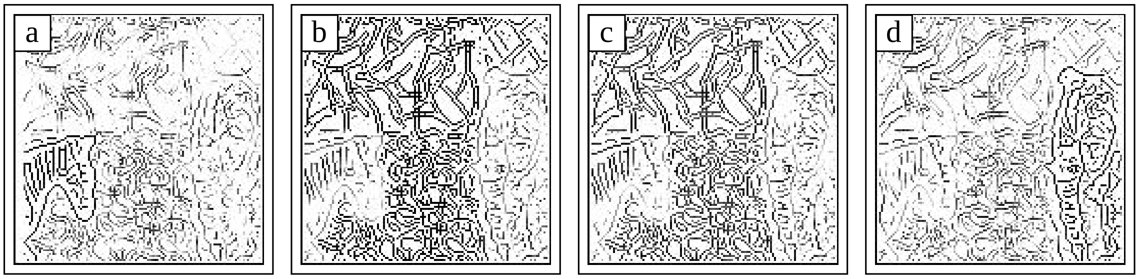

Figure 6: Accordance maps for Frame 15 with respect to the four motion hypotheses as extracted from the image flow histogram shown in Fig. 3. High values are shown dark. **(a)** corresponds to the zebra, **(b)** to the background, **(c)** to the spurious motion hypothesis, and **(d)** to the elephant.

2.1.4 Segmentation

The last stage of the two-frame method finally performs the segmentation. Each edge pixel is classified as belonging to that motion hypothesis for which it has the highest accordance value. One can think of it as a pixel-wise winner-takes-all competition between the accordance maps; see Fig. 7. Notice the increased resolution and reliability as compared to a segmentation result one would obtain based on the image flow of Fig. 2.

2.2 Segmentation on a Sequence of Frames

The reliability of the segmentation can be improved significantly by using a sequence of frames instead of pairs. The following describes a method of associating motion hypotheses between successive pairs of frames and of integrating their accordance maps over time.

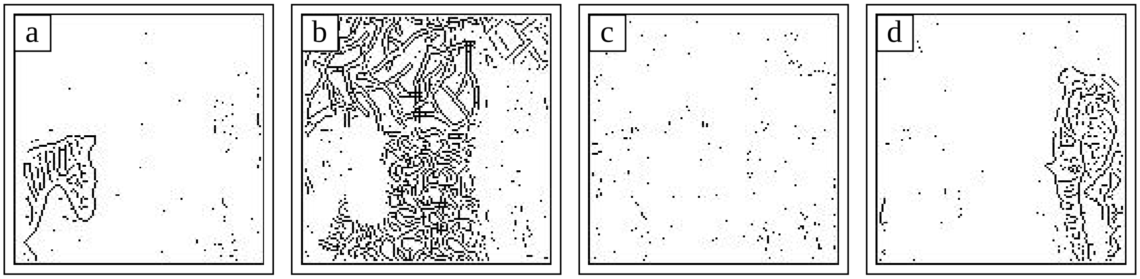

Figure 7: **(a-d)** Segmentation result based on the accordance maps shown in Fig. 6 (a-d), respectively.

2.2.1 Tracking Motion Hypotheses

Consider three successive frames forming two pairs. The segmentation method described above can be applied to each pair independently, but better results can be achieved by integrating the accordances of both pairs. To do this, one has to determine which motion hypothesis of the first pair, frames one and two, correspond to which hypothesis in the second pair, frames two and three. Since the objects may move arbitrarily, e.g. upwards from frame one to frame two and downwards from frame two to frame three, the motion histograms may look very different. The correspondences between the motion hypotheses cannot be found simply based on their similarities.

Instead the motion hypotheses are associated via the spatial domain. Due to the segmentation as described above, a motion hypothesis for the first pair of frames is associated with a set of edge pixels, the segmented object. Generating a motion histogram on the second pair of frames only over these pixels will yield an *object-specific motion histogram* which differs from the complete motion histogram, evaluated on all edge pixels, in a characteristic way; see Fig. 8. In the object-specific motion histogram, all but one local maxima will be suppressed significantly relative to the complete motion histogram. The assumption here is that a region which has moved as a coherent object from the first frame to the second frame will do so from the second to the third frame as well. The unsuppressed local maximum then represents a motion hypothesis which has to be associated with the considered motion hypothesis in the first pair of frames, which serves as a predecessor for the current motion hypothesis; see Fig. 9.

To compare motion hypotheses of the general motion histogram with those from the object-specific motion histogram, the similarity function

$$\mathcal{S}_v(\mathbf{v}_n, \mathbf{v}_{ml}) = \min\left\{\frac{h(\mathbf{v}_n)}{h_m(\mathbf{v}_{ml})}, \frac{h_m(\mathbf{v}_{ml})}{h(\mathbf{v}_n)}\right\} \max\left\{\ 1 - \frac{(\mathbf{v}_n - \mathbf{v}_{ml})^2}{r^2},\ 0\ \right\} \tag{12}$$

is defined, where $\mathbf{v}_n$ and $\mathbf{v}_{ml}$ denote the considered general and object-specific motion hypotheses and $h(\mathbf{v}_n)$ and $h_m(\mathbf{v}_{ml})$ denote the values of the histograms. Two motion hypotheses are similar if the respective local maxima are at the same location and of similar height. Notice that the general histogram and the object-specific histograms are generated for the same pair of frames, such that one may require that the locations of corresponding maxima are close to each other. The locations of the histogram maxima in the preceding pair of frames may be arbitrary. Additional care is taken that a motion hypothesis cannot have more than one predecessor (and successor). If in doubt, the motion hypotheses with a higher maximum in the motion histogram have priority over those with smaller ones. In this way, motion hypotheses can be tracked over sequences of frames and the accordances can be integrated over time. Thus the system is able to track several individual objects over many frames as long as they move in different directions.

2.2.2 Edge Valuation with Integrated Accordances

In the two-frame method, the edge valuation consists of computing the accordance map for each motion hypothesis. In the sequence method, this is supplemented by taking the geometric mean over the accordance as computed for the two-frame method and the accordance of the previous pair of frames, the latter being a result of such an averaging process itself. If the current motion hypothesis has no predecessor or if the accordance value of the predecessor is low, a minimal constant value $\mathcal{A}_\Theta$ is taken instead.

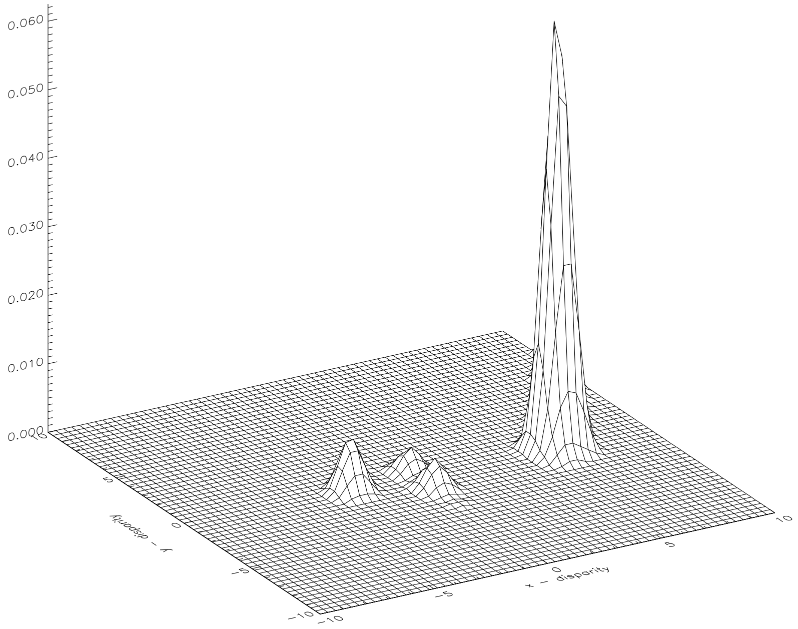

Figure 8: Object-specific image flow histogram for Frames 14 and 15. In contrast to the histogram in Fig. 3, this histogram does not include the flow vectors at all edge pixels. Only those pixels which have been associated with object (a), i.e. the zebra, in the previous pair of frames, Frames 13 and 14, are taken into account. Comparison with Fig. 3 shows that all but the rightmost peak are suppressed. That means that the motion hypothesis of the rightmost peak corresponds to the preceding motion hypothesis which generated object (a). Notice the different range of the z-axis.

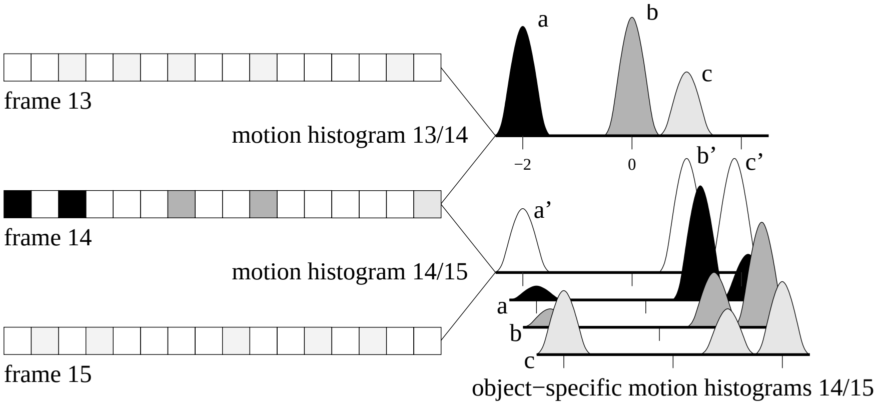

Figure 9: A one-dimensional illustration of motion hypothesis tracking: The motion histogram between Frames 13 and 14 yields three motion hypotheses *a*, *b*, and *c*. The edge pixels of Frame 14 are segmented into corresponding three classes, indicated by different grey values. The motion histogram between Frames 14 and 15 also yields three motion hypotheses *a'*, *b'*, and *c'*. The correspondences between the motion hypotheses cannot be inferred from their similarities, because the object's directions of motion have changed too much. But the object-specific motion histograms give a hint. Histogram *a* contains only those image flow vectors of Frame 14 which belong to edge pixels classified as belonging to object *a*. Its peaks *a'* and *c'* are suppressed while *b'* is not, indicating that *b'* corresponds to *a*. Similarly it can be concluded that *a'* corresponds to *c* and *c'* to *b*.

The integrated accordance is defined as

$$\mathcal{A}(\mathbf{x}, t, \mathbf{v}_n) = \max_{\mathbf{d}} \left\{ \sqrt{\mathcal{S}_d\left(\mathbf{v}_n, \mathbf{d}\right) \mathcal{S}_g\left(\mathbf{g}(\mathbf{x}, t), \mathbf{g}(\mathbf{x} - \mathbf{d}, t-1)\right) \max\left\{\mathcal{A}_\Theta, \mathcal{A}\left(\mathbf{x} - \mathbf{d}, t-1, \mathbf{v}_{n'}\right)\right\}} \right\}, \tag{13}$$

where $\mathbf{v}_{n'}$ refers to the preceding motion hypothesis. By this means, the accordances can be integrated over time and their reliability improves significantly; see Fig. 10.

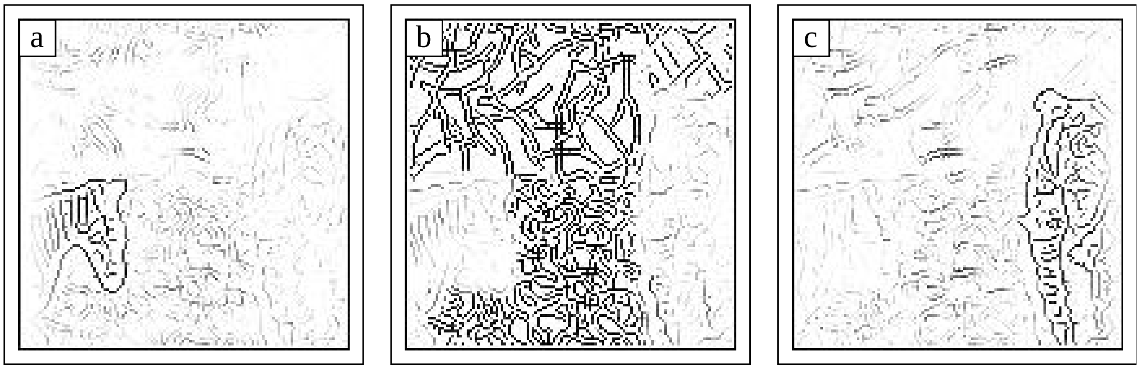

Figure 10: Accordance maps for Frame 15 with respect to the three relevant motion hypotheses, disregarding the spurious one, as extracted from the image flow histogram shown in Fig. 3. These accordance maps have been integrated over time, taking into account the results from the preceding pairs of frames. High values are shown dark. **(a)** corresponds to the zebra, **(b)** to the background, and **(c)** to the elephant. A comparison with Fig. 6 reveals two advantages of the sequence method over the two-frame method. Firstly, the spurious motion hypothesis could be ruled out because it did not have a predecessor. Secondly, the accordances are improved significantly due to the integration over time.

2.2.3 Segmentation

For the sequence method, two types of segmentation are distinguished. The first, called internal segmentation, serves the next pair of frames to generate the object-specific histograms. The second type, called external segmentation, is the actual output result of the system. Both are obtained in the same way as in the two-frame method, with the exception that, for the external segmentation, edge pixels with values which are too low in all accordance maps are disregarded. Another difference is that the internal segmentation takes all motion hypotheses into consideration, while the external segmentation ignores those which have no predecessor. This eliminates spurious motion hypotheses which appear only in a single pair of frames, but it allows consideration of objects newly entering the scene. Fig. 11 shows the segmentation result for the moving-animals sequence with temporal integration.

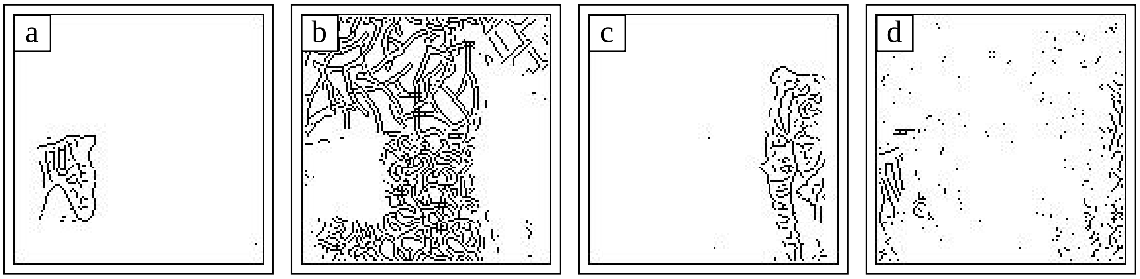

Figure 11: **(a-c)** Segmentation result based on the accordance maps shown in Fig. 10 (a-c), respectively. **(d)** shows the pixels which were not categorized, because their accordance values were too low.

3 Examples

One strength of the system is that it generates motion hypotheses on a coarse level, but segments on a single pixel level. This allows the system to segment even small, disconnected, or openworked objects. This is demonstrated in Fig. 12 and 13. A second strength is that the motions need not be continuous in order to perform the integration over a sequence of frames. Objects may jump back and forth, and the method will still be able to track them. This is demonstrated in Fig. 13, where random shifts of the frames relative to each other have been introduced artificially. The most restrictive assumption of the system is that objects move fronto-parallel. Fig.14 shows an example of how the system behaves if this assumption is not met. A person walks into a room and his swinging hand is not segmented as belonging to the body, because the hand moves too fast relative to the body. An even more extreme example of violating the fronto-parallel-motion assumption is shown in Fig. 15. The plant is rotating in depth such that all leaves move relative to each other. The system treats this inappropriate sequence by breaking it down into parts which can be approximated as moving fronto-parallel. The leaves moving fast to the left are grouped together and those which move slowly are grouped with the background. A similar result was found for objects rotating in the image plane.

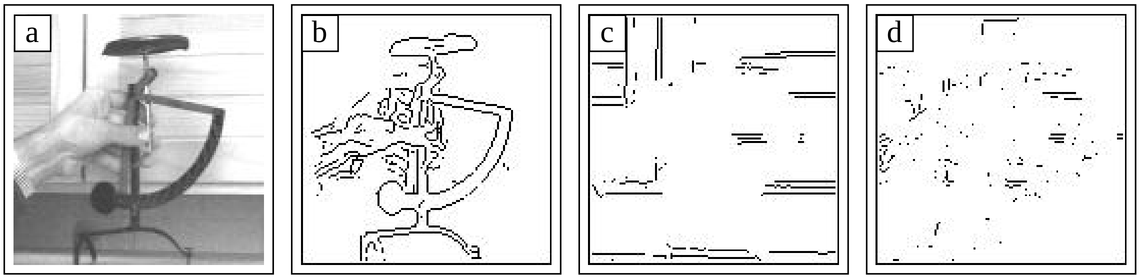

Figure 12: **(a)** The ninth frame of a letter-scale sequence. **(b) (c)** Segmentation result for the letter scale and the background respectively. **(d)** Uncategorized pixels. In this sequence it sometimes happened that a third motion hypothesis had a predecessor, and hence was erroneously interpreted as belonging to a real object.

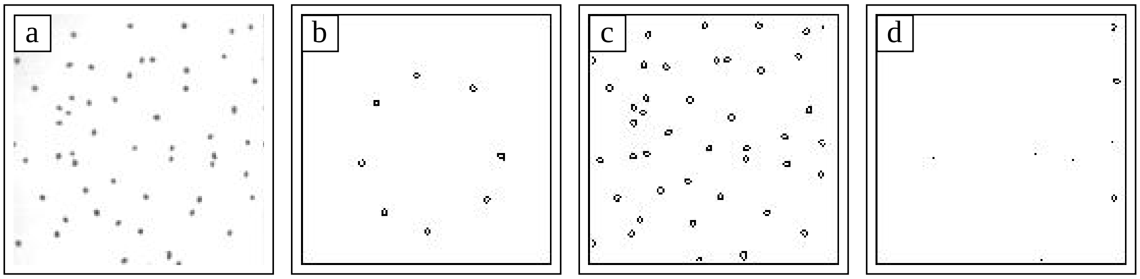

Figure 13: **(a)** The seventh frame of a dot-pattern sequence, showing a stationary circle with eight points and a dotted background moving left. Each frame in the sequence is additionally shifted randomly up to $(\pm 2, \pm 2)$ pixels, resulting in an additional relative displacement of up to ± 4 pixels in each dimension. **(b) (c)** Segmentation result for the circle and the background respectively. **(d)** Uncategorized pixels. One can see three dots entering the image on the right and not yet having accumulated enough evidence for segmentation. For this sequence, even the two-frame method yields a segmentation result of this high quality.

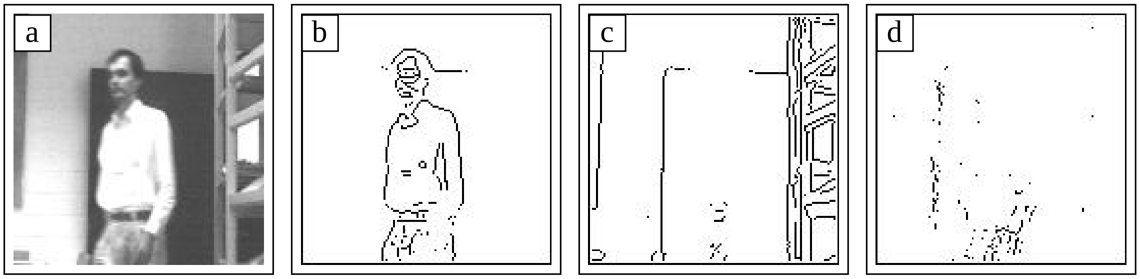

Figure 14: **(a)** The fourth frame of a walking-person sequence. **(b) (c)** Segmentation result for the person and the background respectively. **(d)** Uncategorized pixels. In this frame the hand is not segmented as belonging to the body, due to the swinging of the arm.

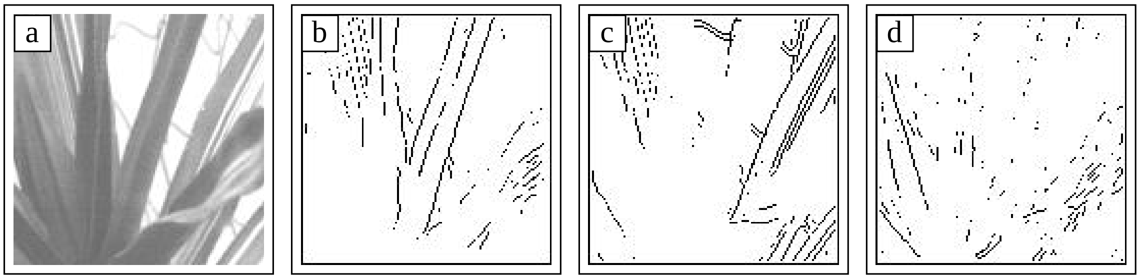

Figure 15: **(a)** The fifth frame of a rotating-plant sequence. **(b) (c)** Segmentation result for the fast- and the slow-moving leaves respectively. **(d)** Uncategorized pixels.

4 Discussion

4.1 Comparison with Other Systems

The system presented here differs significantly from the two dominant classes of systems for segmentation from motion described in the introduction, filter-based and matching-based methods. However, most of the components used in this system are techniques known in the literature. The image flow estimation is adopted from a stereo algorithm [34, 36], and a similar system has been presented elsewhere [37]. Histogram techniques for finding the dominant motions can be found in various systems [38, 39, 40]. Matching feature points for image flow estimation has been used often [10, 11, 12]. Matching of edges, however, seems to be used only rarely [38, 40]. The accordance maps and the representation of the segmentation results are somewhat analogous to the layers used in some systems [7, 18], though there are significant differences. Temporal integration has also been used before, though it is usually done by assuming continuous motion [31, 12]. But there is also an example of temporal integration without continuity assumptions [41]. However, this latter system integrates the gray value images over several frames, while the system presented here integrates segmentation evidence.

What is the original contribution of the system presented here? The key idea is to combine two different representations (and methods) for the same cue to overcome the weaknesses of either of them. Gabor-wavelets allow an accurate and unique estimate of image flow, but only with low spatial resolution. Mallat-wavelets, on the other hand, provide high-resolution information (relative to the frequency band of the edges) but with a lot of ambiguity as to which edges correspond to which edges in the next frame. By combining the two, reliable high-resolution segmentation can be achieved. To my knowledge, this concept has not been put forward before. Systems for segmentation from motion usually rely on one representation only. There are, of course, examples of combining representations of different cues, such as motion and intensity [42, 43], and systems often integrate information over a certain spatial neighborhood to improve image flow estimation. But this is different from combining qualitatively different representations for the same cue.

Most similar to this kind of integrative approach may be the hierarchical systems for image flow estimation integrating over spatial scale or frequency bands [20, 44, 45]. These systems usually estimate image flow and segment boundaries on a coarse scale and then use this information to guide the same process on the next finer scale. However, the representation used on different scales is always the same and its weakness may remain a problem. The image flow algorithm based on Gabor-wavelets used here is such a hierarchical algorithm. It is important to notice that combining Gabor-wavelets with Mallat-wavelets is not hierarchical in this sense but qualitatively different. Gabor- and Mallat-wavelets may look at the same frequency band, in which case the Mallat-wavelets have a much smaller support, or they may have the same support, in which case the frequency band of the Mallat-wavelet is lower. It is this difference in quality and not in quantity that is characteristic for the system presented.

There are two systems that are more closely related to the system presented here and that will now be discussed in greater detail. Huttenlocher et al. [46] developed a system for tracking moving objects which is also based on edges and a translational motion model. It is thus closely related to the latter stages of the system presented here. Each frame is represented as a binary edge image, encoding just for the presence of edges, but neither for their gradient magnitude nor their direction angle. In the first frame, a model of the object to be tracked has to be defined. This is done by manually selecting a region in the image. The model is a binary edge image as well. The model is then matched to the next frame by using a Hausdorff distance, which may be thought of as a fuzzy template matching. This matching accounts only for translation, but it is not restricted with respect to the magnitude of the displacements. An updated model is generated by selecting those edge pixels which are close to edge pixels of the matched model. The system has some robustness with respect to distortions, but if, for instance, the arms of a walking person are swinging too quickly, the arms of the model become truncated. On the other hand, if the background is cluttered, background edges may become integrated into the model. Both the distortion robustness and the sensitivity to cluttered background are determined by the size of the neighborhood of edges which is taken into account in generating the updated model. The system is also able to track several models simultaneously.

The output of this system looks similar to the examples shown in Section 3 (cf., for instance, Fig. 14 (b) with Fig. 3 in reference [46]). However, there are several differences between the two systems. First of all, the system of Huttenlocher et al. [46] is not designed for segmentation but for tracking. It requires an initial definition of a model. However, if the background is stationary, fairly simple motion detection methods can be used to define models automatically. On the other hand, the explicit object model allows the system to track objects even if they disappear temporarily. In that case, the system can continue to match stored models of previous frames until the object reappears. This higher level of model representation is not part of the system presented here. Secondly, in the system of Huttenlocher et al. [46], the displacement of objects between two frames is estimated by matching. This has the advantage of having no restrictions regarding the displacement magnitude, but it has the problem that it is easily misled in the presence of multiple objects of similar shape. In the system presented here, the motion hypotheses are restricted in their magnitude by the low frequency Gabor-wavelets, but, in combination with the hypothesis tracking mechanism, they provide object-specific displacement estimations and can deal with multiple identical objects. Thirdly, the simple binary edge representation results in ambiguities which make the system sensitive to cluttered background and which restricts the applicability to compact objects or simple background, fairly blank or resting. In the latter case, resting edges can be suppressed. A more complete edge representation, such as used here, would reduce the ambiguities and might solve this problem; see, for instance, Fig. 5 (c), where a lot of background edge pixels are present. Fourthly, also contributing to this background problem is the fact that the background is not represented by a model. So, the question in matching a given edge pixel is not whether it better fits a model or the background, but rather whether it fits a model well at all. The latter question is in general harder to decide and requires an appropriate threshold. In the system presented here, the evidence for each motion hypothesis is first accumulated in an accordance map, then the segmentation is done by a maximum detection over all maps, rather than a simple thresholding.

Cumani et al. [40] have developed a system for segmentation from motion which is also based on an edge description and a translational motion model. Contour points are detected as zero crossings of the second directional derivative of the image grey value. The displacement of an edge is found by matching the contour point along a line perpendicular to its orientation, which yields only the motion component in that direction. The translational motion direction of whole objects is then found as peaks in a Hough transform. This is analogous to the motion hypothesis used here. Contour segments are then assigned to that object for which their displacement estimates are consistent. Segments that are either consistent with two objects or with none are not assigned. The number of assigned contour segments can be increased if information on adjacent segments is integrated to yield a unique motion direction. No temporal integration is used to improve the segmentation result. The performance of the system is demonstrated by reconstructing images of the segmented objects from the segmented contour segments.

This system is similar to the system presented here in that edges are segmented according to motion hypotheses. However, the system of Cumani et al. [40] differs in many technical aspects. For instance, the motion hypotheses are derived from a Hough transform based on the contour segments, while in the system presented here the motion hypotheses are derived from a simple histogram derived from image flow computed on the basis of Gabor-wavelets. In this sense, the system of Cumani et al. [40] is based more on algorithmic complexity while the system presented here is based more on a richer representation. The detection and representation of edges is also quite different.

4.2 Limitations and Extensions

The examples of Section 3 show some of the strengths of the system presented here. The combination of Gabor-wavelets with large support and more localized Mallat-wavelets avoids the aperture problem, reduces the correspondence problem, and therefore provides a more reliable segmentation with high resolution. No restrictions are imposed on the number or shape of the objects. This allows good segmentation even of random dot patterns. Temporal integration improves segmentation results, without assuming motion continuity. In these respects, the system is general and performs well.

One may regard the fact that only edges are segmented as a restriction. One may argue that the ultimate goal of segmentation should be partitioning the full field. However, Mallat and Zhong [25] have shown that edge information as used here on several resolution levels is sufficient to represent and reconstruct a whole image. Thus, if required, one could apply the edge segmentation to several levels and reconstruct the segmented objects from those edge images. Cumani et al. [40] demonstrated good reconstruction results for their edge representations. The alternative approach, segmenting image regions directly, has its own problems. First of all, regions cannot be uniquely assigned to objects, which becomes obvious for the gap between two moving objects. Even if they could, e.g. by assigning ambiguous regions to the background, typical image flow algorithms do not provide segmentation evidence along the edges. Image flow discontinuities typically run through homogeneous regions, e.g. Fig. 12 in reference [7]. These systems therefore have a strong tendency not to segment along object boundaries. An exception are the systems which perform an edge-based segmentation on single frames before image flow estimation [6].

Neither the two-frame segmentation method nor the sequence method include object model constraints. Each pixel is segmented independently of the neighboring ones. This makes it possible to segment several small, disconnected, or openworked objects. However, in some cases one wants to impose constraints given by an object model. The accordance maps are probably the best place to introduce such constraints, since they form a concise representation of all low-level evidence which has been accumulated for segmentation, but without making any final decision. One could either enhance the accordance maps, e.g. by applying a median filter on the edge pixels, and still use the simple segmentation rule presented here, or one could replace this local segmentation rule by a more global one, e.g. forcing pixels of the same line segment to be segmented into one class.

Similar considerations hold for temporal integration. No temporal continuity constraints are imposed; objects may jump back and forth. However, a Kalman-filter-type continuity model could easily be integrated into the hypothesis tracking stage or into the computation of the integrated accordance maps.

The most severe limitation of the system presented is the restriction to translational motion. Though the system degrades gracefully, its limitations are clearly visible in Fig. 14 and 15. But one should notice that, for these sequences, the affine motion model, which is the most popular model, would fail as well. Even the 3D-rigid motion model, which could account for the sequence in Fig. 15, would fail for non-rigid objects as frequently encountered in the real world, e.g. Fig. 14. Any given motion model will break down sooner or later. Furthermore, the notion of individual moving objects might be a misconception in any case. This can be illustrated in Fig. 14. One may argue that the hand moves relative to the body of the person and therefore has to be segmented as a separate object. One may also argue that the hand is connected with the body and therefore has to be segmented as belonging to the body. This conflict cannot be resolved within the paradigm of separately moving objects. Both problems, the breakdown of any motion model and the argument of connected vs. unconnected object arrangements, are related and can be resolved with a different segmentation paradigm.

I propose to perform the segmentation locally and integrate the local fragments on a higher level. One could, for example, apply a simple segmentation method to small overlapping image patches. The method could be restricted to pure translational motion and to, for instance, three possible components. Then each patch would be segmented into one, two, or three parts. Due to the overlap between the patches, corresponding parts in neighboring patches will have an overlap and can be joined together. This joining, which may be improved by regularization constraints and with explicit object knowledge, connects regions over longer distances while preserving local segmentation decisions. The result could take the form of a graph, whose nodes represent the edges and whose vertices represent whether edges are connected or not. Thus, the hand could be connected with the body and locally segregated from the body at the same time. This modified segmentation paradigm for edge representations is analogous to image flow algorithms for dense pixel representations that introduce discontinuity lines that are not necessarily closed [47, 48]. In such a local segmentation process, the translational motion model may be the most appropriate one because of its simplicity.

4.3 Conclusion

The system presented here performs segmentation from motion on a sequence of frames. It integrates two different techniques based on Gabor- and Mallat-wavelets to overcome the aperture and the correspondence problems, and it integrates over time for an additional improvement of the segmentation result. Segmentation is only performed on edges and it is argued that edges are the most appropriate representation for segmentation from motion, because they provide enough motion evidence and the complete grey-value distribution could be reconstructed from multi-resolution edge information. Since no object model is used, the system can segment small and openworked objects. The performance of the system is improved by temporal integration, without making any assumption about motion continuity. The motion model used is pure translation and obviously inappropriate for objects in arbitrary motion. However, it is argued that the traditional paradigm for segmentation from motion that strives to segment individual objects globally on the basis of given motion models is inappropriate in any case. This paradigm is assumed here as well, but I advocate performing segmentation locally and integrating the local segmentation decisions on a higher level. The translational motion model may then be most appropriate because of its simplicity, and the system presented here could serve as a local mechanism.

Acknowledgement

Many thanks go to Prof. von der Malsburg for his support and helpful comments and to an anonymous referee for carefully reading the manuscript. This work was supported by grants from the German Federal Ministry for Science and Technology (413-5839-01 IN 101 B9).

References

[1] A. Mitiche and J. K. Aggarwal. Image segmentation by conventional and information-integrating techniques: a synopsis. *Image and Vision Computing*, 3(2):50–62, 1985. 1

[2] V. Cappellini, A. Mecocci, and A. Del Bimbo. Motion analysis and representation in computer vision. *Journal of Circuits, Systems and Computers*, 3(4):797–831, 1993. 1

[3] C. L. Fennema and W. B. Thompson. Velocity determination in scenes containing several moving objects. *Computer Graphics and Image Processing*, 9:301–315, 1979. 1

[4] B. K. P. Horn and B. G. Schunck. Determining optical flow. *Artificial Intelligence*, 17:185–203, 1981. 1, 2

[5] J. R. Bergen, P. J. Burt, R. Hingorani, and S. Peleg. A three-frame algorithm for estimating two-component image motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(9):886–895, 1992. 1

[6] H. Morikawa and H. Harashima. Incremental segmentation of moving pictures: An analysis by synthesis approach. *IEICE Trans. Inf. & Syst.*, E76-D(4):446–453, 1993. 1, 2, 16

[7] J. Y. A. Wang and E. H. Adelson. Representing moving images with layers. *IEEE Transactions on Image Processing*, 3(5):625–638, 1994. 1, 14, 16

[8] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–7104, 1996. 1

[9] M. M. Chang, A. M. Tekalp, and M. I. Sezan. Simultaneous motion estimation and segmentation. *IEEE Transactions on Image Processing*, 6(9):1326–1333, 1997. 1

[10] M. Shah, K. Rangarajan, and P.-S. Tsai. Motion trajectories. *IEEE Transactions on Systems, Man and Cybernetics*, 23(4):1138–1150, 1993. 1, 14

[11] W. B. Thompson, P. Lechleider, and E. R. Stuck. Detecting moving objects using the rigidity constraint. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(2):162–166, 1993. 1, 14

[12] S. M. Smith and J. M. Brady. ASSET-2: Real-time motion segmentation and shape tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):814–820, 1995. 1, 3, 14

[13] M. T. Orchard. Predictive motion-field segmentation for image sequence coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 3(1):54–70, 1993. 2

[14] H. Zheng and S. D. Blostein. Motion-based object segmentation and estimation using the MDL principle. *IEEE Transactions on Image Processing*, 4(9):1223–1235, 1995. 2

[15] G. Healey. Hierarchical segmentation-based approach to motion analysis. *Image and Vision Computing*, 11(9):570–576, 1993. 2

[16] L. Wiskott and C. von der Malsburg. A neural system for the recognition of partially occluded objects in cluttered scenes. *Int. J. of Pattern Recognition and Artificial Intelligence*, 7(4):935–948, 1993. 2

[17] L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997. 2, 4, 5

[18] T. Darrell and A. P. Pentland. Cooperative robust estimation using layers of support. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(5):474–487, 1995. 2, 14

[19] Y. Huang, K. Palaniappan, X. Zhuang, and J. E. Cavanaugh. Optic flow field segmentation and motion estimation using a robust genetic partitioning algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(12):1177–1190, 1995. 2

[20] P. Anandan. A computational framework and an algorithm for the measurement of visual motion. *International Journal of Computer Vision*, 2(3):283–310, 1989. 2, 14

[21] T. Poggio, E. B. Gamble, and J. J. Little. Parallel integration of vision modules. *Science*, 242:436–440, 1988. 2

[22] M.-P. Dubuisson and A. K. Jain. Contour extraction of moving objects in complex outdoor scenes. *Intern. Journal of Computer Vision*, 14(1):83–105, 1995. 2

[23] C. Eckes and J. C. Vorbrüggen. Combining data-driven and model-based cues for segmentation of video sequences. In *Proc. World Congress on Neural Networks*, pages 868–875, San Diego, CA, September 1996. Intern. Neural Network Soc., Lawrence Erlbaum Assoc. Inc. Mahwah, NJ. 2

[24] J. G. Daugman. Complete discrete 2-D Gabor transform by neural networks for image analysis and compression. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 36(7):1169–1179, July 1988. 2, 4, 5

[25] S. Mallat and S. Zhong. Characterization of signals from multiscale edges. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(7):710–732, 1992. 2, 3, 7, 8, 16

[26] J. P. Jones and L. A. Palmer. An evaluation of the two dimensional Gabor filter model of simple receptive fields in cat striate cortex. *J. of Neurophysiology*, 58:1233–1258, 1987. 2, 5

[27] R. L. DeValois and K. K. DeValois. *Spatial Vision.* Oxford Press, 1988. 2, 5

[28] D. C. Burr, M. C. Morrone, and D. Spinelli. Evidence for edge and bar detectors in human vision. *Vision Research*, 29(4):419–431, 1989. 2

[29] E. Börjesson and U. Ahlström. Motion structure in five-dot patterns as a determinant of perceptual grouping. *Perception and Psychophysics*, 53(1):2–12, January 1993. 2

[30] V. S. Ramachandran, S. Cobb, and D. Rogers-Ramachandran. Perception of 3-D structure from motion: The role of velocity gradients and segmentation boundaries. *Perception and Psychophysics*, 44(4):390–393, 1988. 3

[31] P. Bouthemy and E. Francois. Motion segmentation and qualitative dynamic scene analysis from an image sequence. *Intern. Journal of Computer Vision*, 10(2):157–182, 1993. 3, 14

[32] C. Schnörr. Computation of discontinuous optical flow by domain decomposition and shape optimization. *Int. J. of Computer Vision*, 8(2):153–165, 1992. 3

[33] L. Westberg. Hierarchical contour-based segmentation of dynamic scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(9):946–952, September 1992. 3

[34] W. M. Theimer and H. A. Mallot. Phase-based binocular vergence control and depth reconstruction using active vision. *CVGIP: Image Understanding*, 60(3):343–358, November 1994. 5, 14

[35] H.-J. Chen, Y. Shirai, and M. Asada. Detecting multiple rigid image motions from an optical flow field obtained with multi-scale, multi-orientation filters. *IEICE Trans. on Information and Systems*, E76-D(10):1253–1262, 1993. 6

[36] D. J. Fleet and A. D. Jepson. Computation of component image velocity from local phase information. *Intern. Journal of Computer Vision*, 5(1):77–104, 1990. 14

[37] F. Valentinotti, G. Di Caro, and B. Crespi. Real-time parallel computation of disparity and optical flow using phase difference. *Machine Vision and Applications*, 9(3):87–96, 1996. 14

[38] D. T. Lawton. Processing translational motion sequences. *Computer Vision, Graphics and Image Processing*, 22(1):116–144, 1983. 14

[39] B. G. Schunck. Image flow segmentation and estimation by constraint line clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(10):1010–1027, 1989. 14

[40] A. Cumani, A. Guiducci, and P. Grattoni. Image description of dynamic scenes. *Pattern Recognition*, 24(7):661–673, 1991. 14, 15, 16

[41] M. Irani, B. Rousso, and S. Peleg. Computing occluding and transparent motions. *International Journal of Computer Vision*, 12(1):5–16, 1994. 14

[42] W. B. Thompson. Combining motion and contrast for segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(6):543–549, 1980. 14

[43] S.-W. Lee, J. G. Choi, and S.-D. Kim. Scene segmentation using a combined criterion of motion and intensity. *Optical Engineering*, 36(8):2346–2352, 1997. 14

[44] S. H. Hwang and S. U. Lee. A hierarchical optical flow estimation algorithm based on the interlevel motion smoothness constraint. *Pattern Recognition*, 26(6):939–952, 1993. 14

[45] M. R. Luettgen, W. C. Karl, and A. S. Willsky. Efficient multiscale regularization with applications to the computation of optical flow. *IEEE Transactions on Image Processing*, 3(1):41–64, 1994. 14

[46] D. P. Huttenlocher, J. J. Noh, and W. J. Rucklidge. Tracking non-rigid objects in complex scenes. In *Proc. Fourth Intern. Conf. on Computer Vision, Berlin, Germany*, pages 93–101, Los Alamitos, CA, USA, 1993. IEEE Computer Society Press. 15

[47] J. Konrad and E. Dubois. Bayesian estimation of vector motion fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(9):910–927, 1992. 17

[48] F. Heitz and P. Bouthemy. Multimodal motion estimation of discontinuous optical flow using Markov random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(12):1217–1232, 1993. 17